Mapping out MAPK interactors

PhD Thesis

by András Zeke

Principal Investigator: Attila Reményi, PhD

Eötvös Loránd University of Sciences, PhD School of Biology
Structural Biochemistry Programme
László Nyitray, PhD,
university professor

Hungarian Academy of Sciences,
Institute of Enzymology
2015

TABLE OF CONTENTS

INTRODUCTION & OVERVIEW

Brief introduction

Mitogen-activated protein kinases (MAPKs) are ubiquitous components of eukaryotic signalling systems. These proten kinases fulfill a wide range of roles in human physiology, ranging from immunity and DNA damage responses to growth factor signalling and embryonic development. They also have a key pathophysiological role in a number of conditions, including cancer, diabetes, autommunity and several neurodegenerative diseases. Despite the fact that the biology, interactions and structure of these enzymes have been extensively studied for at least two decades, many aspects of MAPK-regulated biochemical systems are still not well understood. We do not know how many substrates or regulators these enzymes have or how these partnerships came into existence. Systematic identification of MAPK-interacting proteins is still a challenging task, both computationally and experimentally. But we shall show that by introducing structurally sound predictions and experimental methods suitable to detect low-affinity interactions, many MAPK-partner proteins become readily identifiable in the human proteome. These novel, often tissue-specific partners (potentially ranging in several hundreds) unravel how this extensive system is “wired” in the human proteome. By studying the evolutionary aspects of protein-protein interactions, we may also begin to understand how such linear motif-based regulatory systems can evolve over time. A deeper analysis of partnerships could also help us to decipher how MAPK-regulated modules are built into proteins; and how they contribute to the pathomechanism of certain diseases.

Historic overview and nomenclature

The first MAPK to be discovered were the mammalian ERK1 and ERK2 proteins. They were originally identified as the key component of cell extracts responsible for the phosphorylation of microtubule associated protein 2 (MAP2): earning the name "**MAP kinase**". But researchers in 1990 realized that the 42 to 44 kDa extracellularly regulated kinases (ERK1 and ERK2) phosphorylated in serum-stimulated cell extracts were identical to the microtubule associated protein 2 kinase (MAPK) described a few years before.[1–3] To reflect this identification, the earlier name was updated to become "**Mitogen-activated protein kinase**" while keeping the old acronym. It was soon realized that ERK1 and ERK2 are critical components of growth factor pathways, acting downstream of many proto-oncogens: tyrosine kinase receptors, Ras and Raf proteins, justifying their name. Following ERK1 and ERK2 in the subsequent years, several more mammalian MAPKs were described. However, these appeared to be activated by cellular stress stimuli, rather than growth factors. The first c-Jun N-terminal kinase (JNK) was discovered in 1993 as the kinase responsible for the phosphorylation and activation of the transcription factor c-Jun.[4] The first stress-responsive 38-kDa MAPK, p38α was identified in 1993 and cloned in 1995.[5,6] Later it turned out that each of these proteins also have other close relatives in the human proteome. Therefore these new kinases were named by adding either numbers or Greek letters to the previously-established names (JNK1, JNK2, p38α, p38β, etc.).[7]

The nomenclature of MAPKs is unfortunately complex, and not very consistent. Many old synonyms are still occasionally in use and generally result in confusion. As with other protein families, several newly described MAPK genes and their products received names in a rather unsystematic fashion. Initially, JNK1 was also called SAPK (stress-activated protein kinase), but the name "SAPK2" was never used for JNK2, being reserved for p38α instead.[7] At the earliest stage, the term "ERK" (extracellularly regulated kinase) was also understood as a synonym for "MAPK" in general. This resulted in a number of novel human MAPKs being named ERKs, like ERK3, ERK4 and ERK5, lacking any close relationsip to the previously described ERK1 or ERK2.[8] To add to the confusion, p38γ was initially described as ERK6, while the names ERK7 and ERK8 were used for distinct mouse and human proteins, that later turned out to be homologous to each other.[9,10] By that time, several MAPKs were described from non-mammalian organisms as well, including the fruitfly *Drosophila melanogaster*, the roundworm *Caenorhabditis elegans*, the fungi *Saccharomyces cerevisiae* and *Schisosaccharomyces pombe* and even higher plants, like *Arabidopsis thaliana*. The flurry of names

given to genes and proteins with complete disregard to sequence similarities or evolutionary homologies quickly made the situation even worse.

To remedy the situation, in 2004 the Gene Nomenclature Committee of the Human Genome Project (HUGO) decided to let go of all earlier names like ERK, JNK or p38. Instead, it was suggested that all human MAPKs should receive the name “MAPK” and a systematic numbering. Thus ERK2 is now termed “MAPK1”, ERK1 as MAPK3, etc. In theory, this should have solved problem of multiple names. However, the new “standard nomenclature” also incurred a number of errors. The names “MAPK2” and “MAPK5” were initially falsely assigned (thus they are no longer used for any extant human MAPK). On the other hand, it was not realized that the Nemo-like kinase (NLK) protein is also a member of the MAPK family: hence it did not receive a proper MAPK code. Due to the serious inconsistencies of the “official nomenclature”, the old names for these kinases are still in use. Because these ones are the most widely used, in the current work I will also refer to all human MAPKs with their traditional name (as laid down by Widmann et al, 1999).[11] Table 1 indicates the synomyms and official symbols for each human MAPKs. Table 2 also gives the evolutionary relationships between MAPKs of different organisms, to make the reading and interpretation of cited literature easier.

As our knowledge on MAPKs expanded, it became increasingly clear that the name “**Mitogen activated protein kinase**” is a misnomer. The role of mammalian ERK1 and ERK2 as mediators of growth factor responses is not a universal, but a highly specialized role. Instead, the absolute majority of eukaryotic MAPKs seems to play a role in stress responses. Osmotic stress, oxidative stress or immunological challenges are just a few stimuli that can elicit activation of the majority of human MAPKs, especially the JNK and p38 kinases.[12–15] On the other hand, it is also true that MAPKs do influence development and morphogenesis, not just in multicellular animals, but also in the collective amoeba *Dictyostelium discoideum* and higher plants, such as *Arabidopsis thaliana*.[16,17]

Classical MAPKs (MAPKs *sensu stricto*)					
#	Commonly used name	Systematic name	Other synonyms (mostly obsolete)	Gene location (human)	UniProt accession ID
1	**ERK1**	MAPK3	p44mapk	16p11.2	P27361
2	**ERK2**	MAPK1	ERK, p42mapk	22q11.2	P28482
3	**ERK5**	MAPK7	BMK1	17p11.2	Q13164
4	**JNK1**	MAPK8	JNK, SAPK, SAPK1c	10q11	P45983
5	**JNK2**	MAPK9	SAPK1a, p54a	5q35	P45984
6	**JNK3**	MAPK10	SAPK1b, p54b	4q22-q23	P53779
7	**p38α**	MAPK14	p38, CSBP1, CSBP2, SAPK2a	6p21.3-p21.2	Q16539
8	**p38β**	MAPK11	SAPK2b	22q13.33	Q15759
9	**p38γ**	MAPK12	ERK6, SAPK3	22q13.3	P53778
10	**p38δ**	MAPK13	SAPK4	6p21	O15264
Atypical MAPKs (MAPKs *sensu lato*)					
#	Commonly used name	Systematic name	Other synonyms (mostly obsolete)	Gene location (human)	UniProt accession ID
11	**ERK3**	MAPK6	p97mapk	15q21	Q16659
12	**ERK4**	MAPK4	-	18q21.1	P31152
13	**ERK7**	MAPK15	ERK8	8q24.3	Q8TD08
14	**NLK**	n/a	-	17q11.2	Q9UBE8

Table 1: Names and synonyms of human mitogen-activated protein kinases (MAPKs)

When speaking about MAPKs, researchers usually refer to the group of kinases more explicitly termed as **classical MAPKs**. These are the “typical” mitogen-activated protein kinases, united by a number of features. They have characteristic double phosphorylation sites (threonine-x-tyrosine) on their activation loop, participate in well-known, three-tiered kinase cascades, and have two well-conserved aspartates at the so-called CD-region responsible for substrate recruitment. There are 10 classical MAPKs in humans; These belong to four, reasonably well characterized groups: ERK1/2, JNKs (JNK1, JNK2, JNK3), p38s (p38α, p38β, p38γ, p38δ) and the orphan ERK5, forming a group of its own

Detailed research on the human proteome made it clear that the MAPK family is much broader than traditionally believed. There are several kinases that have a clear sequence and structural homology to

classical MAPKs, yet different in a number of ways. These divergent members of the MAPK family are more commonly called as **atypical MAPKs**. Many atypical MAPK have only one phosphorylation site on their activation loop, are activated by different kinases than classical ones and lack one or both the conserved aspartic amino acids that recruit partners of classical MAPKs. The human proteome contains four atypical MAPKs: ERK3, ERK4 (which are closely related to each other) as well as the more distant ERK7 and NLK. It is still an open question if atypical MAPKs form a single uniform group or not. The identification of atypical MAPKs in distant eukaryotes, like *Dictyostelium discoideum*, *Giardia lamblia* or *Plasmodium falciparum*, with similarity to the mammalian ERK7, hints at their ancient origin. However, secondary emergence from classical MAPKs cannot be discarded for at least some members. In animals this question cannot be solved unambiguously, while higher plants do have certain atypical MAPKs (e.g. Arabidopsis thaliana group C MAPKs), where loss of the canonical features is possibly secondary. Unfortunately, neither human, nor any other atypical MAPKs have been characterized extensively. Thus our current knowledge on pathways in which atypical MAPKs might participate, is very limited .

<table>
<tr><th colspan="4">Homologous MAPKs of genetic model organisms</th></tr>
<tr><th>MAPK subfamily</th><th>Homo sapiens & Mus musculus</th><th>Drosophila melanogaster</th><th>Caenorhabditis elegans</th></tr>
<tr><td rowspan="2">ERK1/2</td><td>ERK1</td><td rowspan="2">Rolled</td><td rowspan="2">mpk-1</td></tr>
<tr><td>ERK2</td></tr>
<tr><td>ERK5</td><td>ERK5</td><td>N/A</td><td>N/A</td></tr>
<tr><td rowspan="3">JNKs</td><td>JNK1</td><td rowspan="3">Basket</td><td rowspan="3">jnk-1</td></tr>
<tr><td>JNK2</td></tr>
<tr><td>JNK3</td></tr>
<tr><td rowspan="4">p38 kinases (*)</td><td>p38α</td><td rowspan="2">p38a</td><td>pmk-1</td></tr>
<tr><td>p38β</td><td rowspan="2">p38b</td><td>pmk-2</td></tr>
<tr><td>p38γ</td><td rowspan="2">p38c</td><td rowspan="2">pmk-3</td></tr>
<tr><td>p38δ</td></tr>
<tr><td rowspan="2">ERK3/4</td><td>ERK3</td><td rowspan="2">N/A</td><td rowspan="2">N/A</td></tr>
<tr><td>ERK4</td></tr>
<tr><td>ERK7</td><td>ERK7 (ERK8)</td><td>Erk7</td><td>C05D10.2</td></tr>
<tr><td>NLK</td><td>NLK</td><td>Nemo</td><td>lit-1</td></tr>
</table>

Table 2: Homologous MAPKs in the most important model animals (homologies among p38 kinases are uncertain, due to unexpectedly high divergence of p38c and pmk-3 genes)*

Control of MAPK activity: the multi-tiered MAPK pathways

Mitogen-activated protein kinases do not form a signalling system on their own. They are part of a larger regulatory network that controls the activity of MAPKs themselves. Classical MAPKs are typically part of three-tiered "**MAPK pathways**". Although their activation does not truely follow exponential kinetics, these linear pathways are still often referred to as cascades. Elements of the upmost tier are called MAP kinase kinase kinases or **MAP3Ks**. These are allosterically-controlled enzymes, capable of autophosphorylation and autoactivation. MAP3Ks are also responsible for the phosphorylation and activation of the middle tier kinases, called MAP kinase kinases or **MAP2Ks**. The latter kinases can only be activated by phosphorylation; but in turn they phosphorylate the MAPKs, thereby activating them as well. The lowest tier of kinases, MAPKs are the ones responsible for the phosphorylation of myriads of effector proteins. Since MAP2Ks are highly specific in their interactions, the MAPK cascades are rather well-insulated and linear pathways. (This does not mean, that there would be no feedback mechanisms from the lowest tier towards the upper ones).

While the seven human MAP2Ks form a single kinase family on their own, descended from a common ancestor (also known as the STE7 family, a branch of the greater STE kinase group), the upmost tier MAP3Ks are extremely diverse and do not even belong to a single superfamily. They include several members of both the **STE** kinase group (named after the yeast Ste7, Ste11 and Ste20 kinases) and the distant **TKL** (tyrosine-kinase like) group.[18] The former group includes MEKK1, MEKK2, MEKK3, MEKK4, ASK1, ASK2, MEKK15, Tpl/Cot as well as the TAO1 and TAO2 kinases among many others; while the latter contains all Raf kinases (A-Raf, B-Raf, c-Raf, KSR1, KSR2), mixed lineage kinases (MLK1, MLK2, MLK3, DLK, ZAK), the solitary TAK1 and their more distant relatives. In some MAPK pathways, additional kinases are also implicated, aiding the activation of certain MAP3Ks. Called "MAPKKKKs" or **MAP4Ks**, their role in the MAPK pathways is usually non-critical. All MAP4Ks (p21-activated kinases: PAKs, germinal center kinases:GCKs and their relatives) also belong to the STE kinase group, thus relatively closely related to many MAP3Ks and all MAP2Ks. The STE kinase group is notable for its members usually being incorporated into multi-tiered kinase pathways (including not only the MAPK, but also the Hippo/LATS, NDR, AMPK and WNK pathways), due to their abilities to dimerize, autoactivate as well as interact with and phosphorylate different kinase domains.

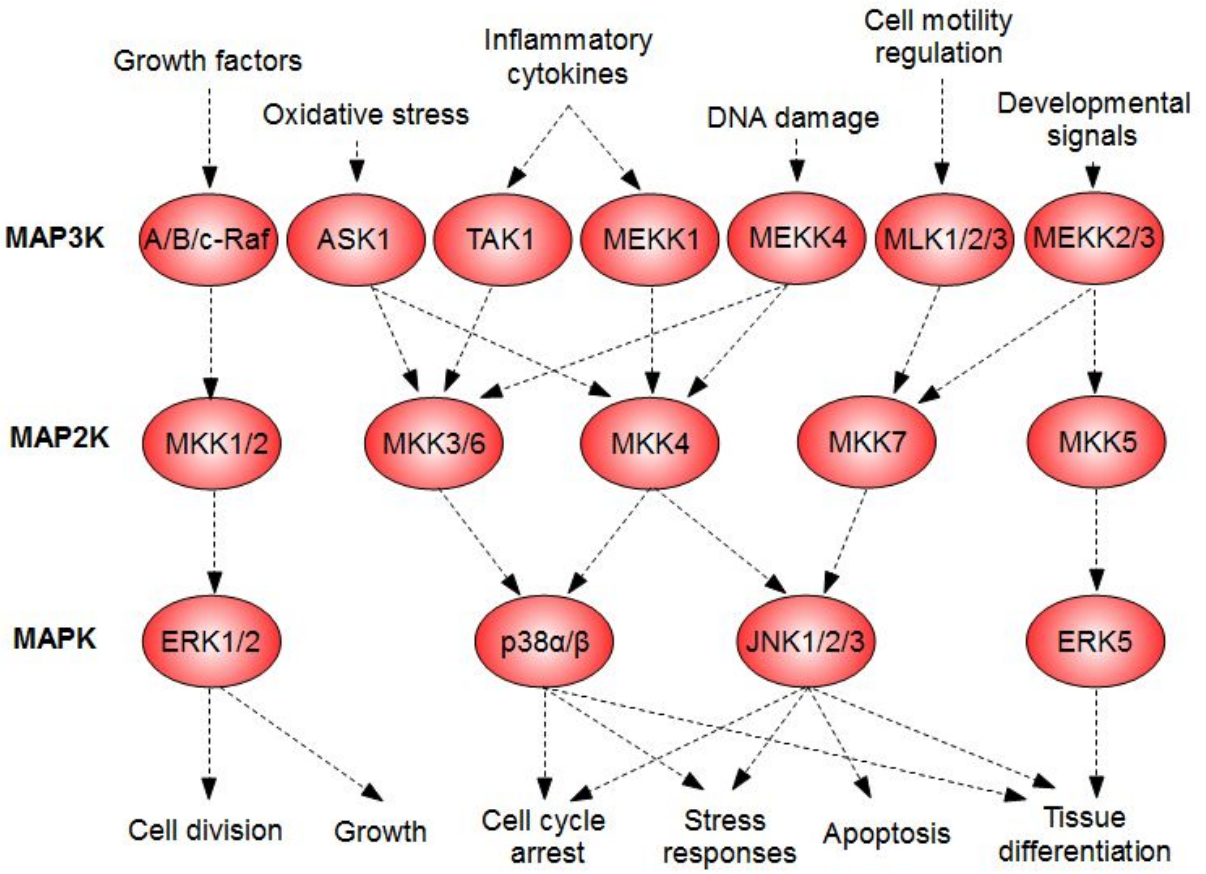


Figure 1: Illustration of the architecture and complexity of human MAPK pathways.

In many MAPK pathways, regardless of the identity of the MAP3K component, the **activation of the pathway** follows the same, strict hierarchy of events:[19] First, external ligads and signals trigger the relief of MAP3K autoinhibition, that liberates their kinase domains. At the next step, the kinase domains spontaneously form dimers (in some cases, also aided by other domains or protein partners). All the MAP3K dimers with a known structure are formed in a symmetric, "back-to back" or "side-to-side" manner that activates the kinase domains through allostery.[20] At this stage, the activity of the enzyme is still low and unstable. As the last step, to achieve full activity, the kinase dimers need to phosphorylate each other on their activation loops. Due to the way the dimers are formed, this step is unlikely to happen within the allosteric kinase dimer itself (and might necessitate the formation of transient, higher-order oligomers). Nevertheless, once the activation loops are phosphorylated (or carry appropriate phospho-mimic mutations, as in the case of oncogenic B-Raf), the MAP3 kinase possesses full activity, no longer dependent on dimerization (and sometimes even loosing sensitivity to the presence of autoinhibitory domains). The phosphorylation of its substrates, most critically the MAP2

kinases takes place once the MAP3Ks are fully activated by autophosphorylation. MAP2Ks contain little more than a kinase domain and an N-terminal MAPK docking motif. Related to STE-type MAP3Ks by structure, MAP2Ks are physiologically devoid of autoactivation capacity: Nevertheless, there are oncogenic mutations in MEK1 that can still "recall" the long-lost ability of this MAP2K to become active on its own.[21] They are very poor enzymes on generic substrates, specializing instead to fit into the kinase domains of MAP3Ks (receiving phosphorylation) and MAPKs (providing phosphorylation) equally effectively. Unfortunately, little is known of the processes leading to inactivation of MAP3Ks and MAP2Ks, though certain kinases, phosphatases and ubiquitin ligases are thought to play a key role.

Major human MAPK pathways

The **ERK1/2 system** is probably the best known human MAPK pathway. This pathway directly couples to growth factor receptors: Dimerization and autophosphorylation of receptor tyrosine kinases allows binding of adaptors (like Grb2) that recruit GTPase exchange factors to the membrane, specific for the small G-protein Ras. The majority of growth factor signalling events are directed by Ras-family GTPases (either H-Ras, K-Ras or N-Ras). Activation and autophosphorylation of Raf (Rapidly accelerated fibrosarcoma) kinases also strictly depend on the presence of GTP-bound Ras proteins. These kinases form the upmost tier of the associated MAPK module. The human genome encodes three different Raf kinases, the prototypical c-Raf, B-Raf (which is a critical proto-oncogen) and the less-studied A-Raf. All known Raf kinases phosphorylate the closely related MEK1 or MEK2 kinases, which in turn phosphorylate and activate ERK1 and ERK2.[22] Once ERK1 or ERK2 has been activated, they phosphorylate a number of proteins in the cytoplasm. However, they are also capable to translocate into the nucleus, activating several transcription factors through phosphorylation. The ERK1/2 pathway is also regulated by a number of non-enzymatic components. Kinase Suppressor of Ras or KSR proteins (which are technically Raf-type kinases, but with very low intrinsic activity) aid the activation of c-Raf (or, to a lesser extent, B-Raf) through heterodimerization and allostery. In addition, KSR1 and KSR2 also provide a point where feedback phosphorylation by ERK1 or ERK2 can negatively regulate pathway activity, as they can be directly regulated by ERK phosphorylation.[23] According to our current knowledge, the ERK1/2 pathway is well insulated from other MAPK systems. MEK1 (MKK1) and MEK2 (MKK2) possess high similarity and have largely redundant roles, as do ERK1 and ERK2. Being responsible for guiding proliferation and cell division, the ERK1/2 pathway is

heavily involved in tumorigenesis. Alteration of receptor tyrosine kinases, all Ras proteins or B-Raf are one of the commonest mutations found in a wide variety of cancers. Inhibitors of B-Raf or MEK1/2 are part of targeted anticancer therapies, and ERK1/2 inhibitors are also in development for the same pharmaceutical purpose.[24]

Unlike the fairly simple ERK1/2 pathway, human **p38 and JNK modules** are much more complicated. Most of the upstream elements are shared between the later two MAPK pathways. They include MEKK1 (MEK/ERK kinase kinase 1), the rather dissimilar MEKK4 (MEK/ERK kinase kinase 4), ASK1 (Apoptosis-stimulating kinase 1), TAK1 (TGF-beta activated kinase) as well as the three MLK (Mixed Lineage Kinase) and DLK (Dual Leucine zipper Kinase) protein kinases. These MAP3Ks have diverse architectures, sharing little homology outside their kinase domains. The structural diversity is explained by the large number of different stress stimuli JNK and p38 systems are activated by. They include the TNFα receptor complex (by TAK1, regulated though non-degradative ubiquitinylation), DNA damage (by MEKK4, activated by GADD45), oxidative stress (by ASK1, regulated by thioredoxine) and cytoskeletal regulation (by MLKs, resposive to Rho-family GTPases).[25–28] All these MAP3 kinases phosphorylate a shared set of MAP2Ks, including MKK3, MKK4, MKK6 and MKK7. Apart from MKK7, which is fairly selectively recruited to mixed lineage kinases by the virtue of a coiled-coil interaction, most MAP3Ks activate multiple different MAP2Ks. Out of the latter, the kinase MKK6 has the highest catalytic activity, both in vitro and in vivo. Together with MKK3, they are mostly responsible for the phosphorylation of p38 kinases, while MKK7 preferentially phosphorylates JNKs. MKK4, on the other hand is truely promiscuous, activating both p38 and JNK kinases to a comparable extent in vitro.[29,30] Unlike the rigid specificity of MEK1 and MEK2 for ERK1/2 (which is “hard-wired” in their kinase domains), the selectivity of MKK6 and MKK4 between JNK or p38 (or the lack thereof) appears to be “soft” and mostly determined by their docking motifs only.[31]

Human genomes encode three, slightly different JNK kinases: JNK1 and JNK2 are ubiquitously expressed, while JNK3 is largely restricted to the central nervous system. c-Jun N-terminal kinases in general, are associated with cell death and apoptosis, especially when overexpressed or activated at a supra-physiological level. On the other hand, they are also critical for physiological regulation of a number of tissues, especially in the central nervous system, where their expression is highest. MKK4 and MKK7 both play an important role in JNK activation. Similarly to other MAPKs, JNKs can also enter the nucleus; thus they are able to phosphorylate nuclear as well as cytoplasmic targets. p38 kinases are even more diverse than JNKs. p38α and p38β are widely expressed, with overlapping roles

in stress responses, inflammation and muscle (both striated muscle and smooth muscle cell) development. On the other hand, p38γ (mostly in striated muscle) and p38δ (skin, digestive tract) have more restricted expression patterns.[32,33] The latter two minor p38 subtypes also differ from p38α and p38β by a number of peculiar structural features. While the physiological targets of p38δ are largely unknown, p38γ seems to have evolved a PDZ domain binding motif, allowing this kinase to selectively phosphorylate components of the neuromuscular junction.[34] Regardless of the structural and functional differences, most p38s are phosphorylated by MKK6 and MKK3 as well as MKK4.[30]

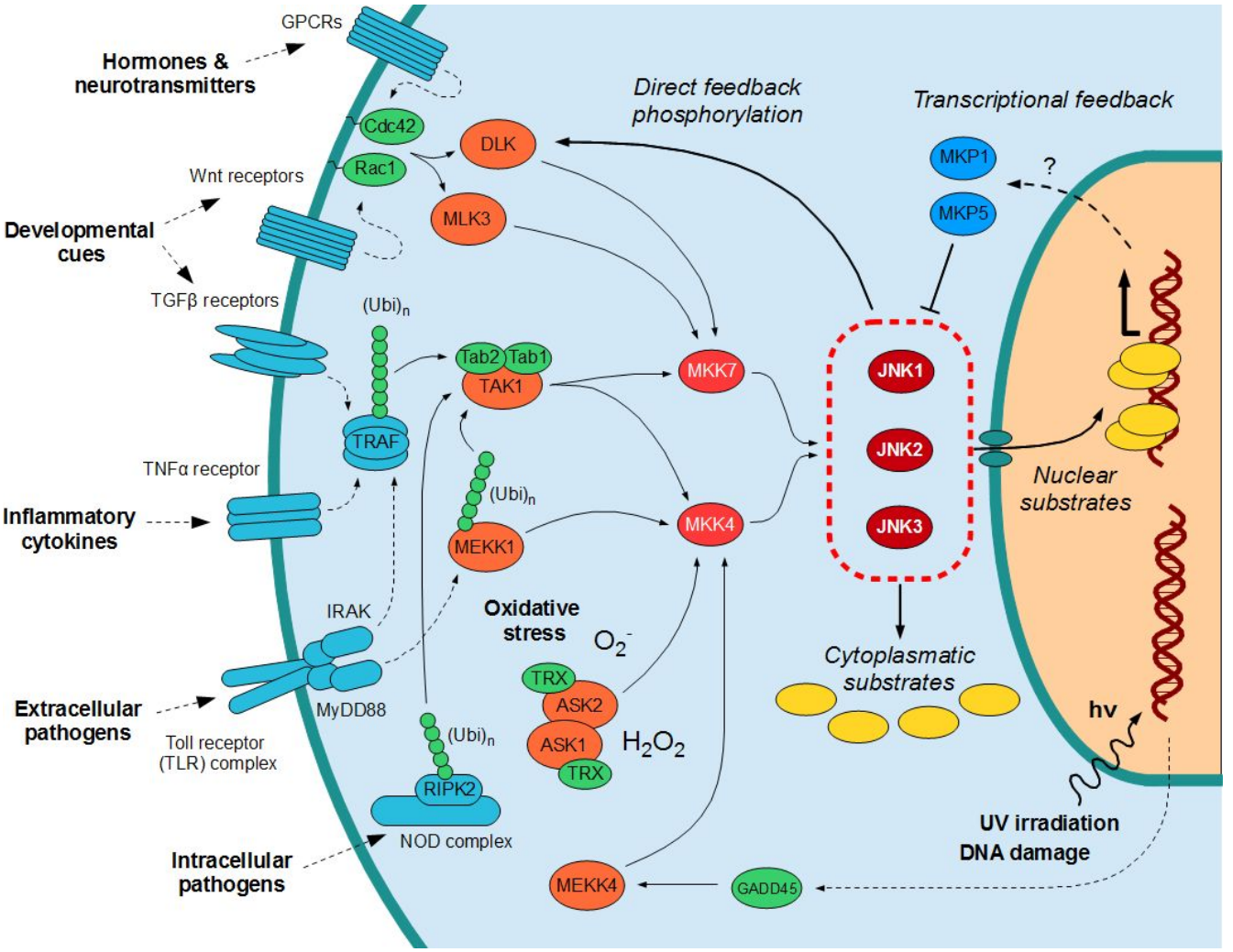


Figure 2: Brief overview of the JNK pathways, activated by diverse extracellular and intracellular stimuli, at the MAP3K level, converging on the three JNK kinases, phosphorylating cytoplasmic & nuclear downstream targets as well as engaging in various feedback circuits.

In addition to their roles in differentiated tissues, both JNK and p38 pathways are critical for embryonic development: although single-knockout MKK6-/- and MKK3-/- animals show no gross abnormalities, the double knockouts (lacking p38 activity) die at an early embryonic stage due to multiplex organ development failures, including impaired haematopoesis, vascular development, neurogenesis, etc.[30] Ablation of MKK4 or MKK7 activity is already lethal on their own: MKK4-/- mice embryos die of impaired liver development, while MKK7-/- mice embryos show a multitude of organ development failures, including impaired liver formation, incomplete neural tube closure and gross defects in brain development.[35–37] Knockout of individual JNK genes usually produces much milder effects, but the double JNK1/JNK2 knockouts already have severe brain malformations, leading to early death after birth.[38]

In contrast to the extensively studied ERK1/2, JNK and p38 pathways, relatively little is known of the fourth major type of classical MAPKs, represented by the single **ERK5** protein. Thanks to its long, presumably disordered C-terminal extension, ERK5 has a much higher molecular weight than the previous proteins, earning the name Big Map Kinase (BMK). Nevertheless, its activation mechanism and substrate recognition are very similar to the former ones. On the other hand, ERK5 is part of an entirely separate MAPK pathway, due to its selective activation by MKK5. MKK5 is unusual in the sense that it contains an N-terminal Phox-Bem1 (PB1) domain in addition to its kinase domain.
At the very top of ERK5 pathway, two MAP3 kinases are responsible for its activation: MEKK2 and the structurally very similar MEKK3.[39] The PB1 domains of MEKK2 and MEKK3 are crucial for the recruitment of MKK5 (through a PB1-PB1 heterodimerization with MKK5) for phosphorylation.[40] As for the MKK5-ERK5 interaction, the PB1 domain of MKK5 also plays a role here, even if this interaction is primarily driven by a docking motif - kinase binding, similarly to other MAP2K-MAPK complexes. While little is known of the molecular mechanisms governing MEKK2 activity (where TNFα receptors are implicated), MEKK3 seems to be directly regulated by the so-called CCM complex. The three CCM proteins, CCM1/Krit1, CCM2 and CCM3 were named for the human disease associated with their mutations, *cerebral cavernous malformations*.[41] The disease is primarily characterized by abnormal endothel formation, and consequential dilation of malformed vessels. Knockout of MEKK3 is embryonic lethal due to severe disruption of heart and blood vessel development, implying critical roles in vasculogenesis - likely connected to those of the CCM complex.[42,43]

Up to date, very few ERK5 substrates were described. The only well-characterized targets are MEF2A and MEF2C. Phosphorylation of MEF2 proteins by ERK5 is important for myogenesis (including the heart and vessel walls), a role strongly overlapping with p38 kinases.[44] Since this pathway (together with all its components) was secondarily deleted in protostomes, genetic experiments on *Drosophila* or *Caenorhabditis* could not have shed light on its function. However, knockout mice models suggest that the ERK5 pathway is essential for viability in mammals: malformations of the primitive heart tube in the ERK5-/- mice cause early embryonic lethality.[45] From experiments on cell cultures we know that ERK5 plays a key role in cardiomyocyte development, vasculogenesis and endothel differentiation in general, but it is also implicated in neurogenesis.[46,47] Interestingly, conditional knockout of the ERK5 gene is also lethal in adults: Shortly after ablation of ERK5 activity, animals die in generalized edema due to widespread disruption of endothelial barriers.[48]

There are several more human proteins that belong to the broader MAPK family. Most of these **atypical MAPKs** are poorly characterized, and we know little of their substrates, let alone mechanisms governing their activation.[49] The **ERK3** and **ERK4** - two closely related proteins - appear to be essential for cell viability and cell cycle regulation. Their only substrate known up to date is the protein kinase MAPKAPK5 (also known as PRAK).[50] These two kinases appear to be activated by phosphorylation, just like classical MAPKs do. The activation loops of ERK3 and ERK4 bears a SEG motif in contrast to the TxY motifs of classical MAPKs. Theferore both their activation and substrate preference is expected to be different. Very recently, the PAK kinases (also implicated as cross-regulating the ERK pathway at a MAP4K level) were detected as being able to activate ERK3 and ERK4.[51] Thus it seems that unlike classical MAPK pathways, at least some atypical MAPKs have a simpler, more ancient, two-tiered system (MAP2Ks are thought to be unable to phosphorylate atypical MAPKs). A different, somewhat less atypical member of the family, **ERK7** bears the same dual phosphorylation motif on its activation loop (TEY) as classical MAPKs do. In sharp contrast, the Nemo-like kinase (**NLK**) has a TQE motif, again hinting at severe deviation from the usual model of MAPK activation. While we know almost nothing of the physiology of ERK7 (including its regulation), NLK appears to form part of inflammatory and/or developmental (especially the Wnt) pathways. It is not impossible that NLK is directly regulated by the MAP3K TAK1, providing another example for two-tiered pathways.[52] However, NLK is also capable of dimerization and autoactivation, which might have a true physiological role in its regulation.[53] The mammalian NLK kinase has an important role in the Wnt pathway; as a consequence, knockout animals suffer from severe growth retardation and show neurological and haematopoetic abnormalities.[54]

MAPK pathways in fungi, plants and other eukaryotes

In addition to mammalian pathways, MAPKs from **yeast** are also well-explored. Although we know substantially less about filamenous fungi, the signalling systems of both the budding yeast *Saccharomyces cerevisiae* and the fission yeast *Schisosaccharomyces pombe* (that are not closely related to each other) utilize multiple different MAPK enzymes. In budding yeast, the **Fus3** kinase is responsible for mediating cellular responses to the **mating pheromone** α-factor, sensed by a heterotrimeric G-protein coupled seven transmembrane (GPCR) receptor. This includes cell cycle arrest and directed growth of mating processes, to allow fusion of haploid cells. As in mammals, the Fus3 pathway is a three-tiered system, consisting of a MAP3K (Ste11), a MAP2K (Ste7) and a MAPK (Fus3) component.[55] In addition, an auxiliary MAP4K enzyme (Ste20) and non-enzymatic components, such as the scaffold protein Ste5 are also required for its proper function. The latter protein has multiple roles, associating with heterotrimeric G-proteins, Ste11, Ste7 and even with Fus3 itself, with mutually non-exclusive interactions.[56] The association of Ste5 (through its von Willebrand domain) with Ste7 is especially important, as this allows for the selective activation of Fus3 in response to pheromone stimulation.[57] This would otherwise not be possible, as Ste7 can also phosphorylate a closely related MAPK, Kss1.[58] Downstream of Fus3, we can find a number of substrates responsible for mediating its actions (such as the E3 ubiquitin ligase Far1, controlling the cell cycle arrest).

Budding yeast also possesses another pathway, sharing many components with the previous one. This is the **filamentous growth pathway**, activated by the lack of critical nutrients (especially the lack of a nitrogen source). The MAP4K (Ste20), MAP3K (Ste11) and MAP2K (Ste7) components are exactly the same as in the mating pathway, but its downstream MAPK target is different: **Kss1** is the kinase preferentially activated by Ste7 in the absence of Ste5.[59,60] The "filamentous growth pathway" received its name for eliciting a radical phenotypical shift: Rounded yeast cells that were previously behaving as a unicellular organism, form long, multicellular, filamentous aggregates (pseudohyphae), roughly similar to the normal vegetative morphology (hyphae) shown by most filamentous fungi. Once active, Kss1 also phosphorylates diverse targets, a number of which are shared with Fus3 (such as the transcription factors Dig1 and Dig2).[61] Fus3 and Kss1 are closely related MAPK paralogs, product of a relatively recent gene duplication in the yeast lineage. Other fungal species, such as *Schisosaccharomyces pombe* have a single pathway in their place (consisting of the MAP3K Byr2, the MAP2K Byr1 and the MAPK Spk1), controlling conjugation and sporulation.[62] As for their broader

relationships, Fus3, Kss1 and Spk1 are all orthologs of mammalian ERK1 and ERK2 kinases.[62] Note that these distant relatives are also required for directing vegetative growth in the diploid phase as well as being involved in meiosis (where ERK1/2 activity is controlled by the unusual MAP3K Mos).[63]

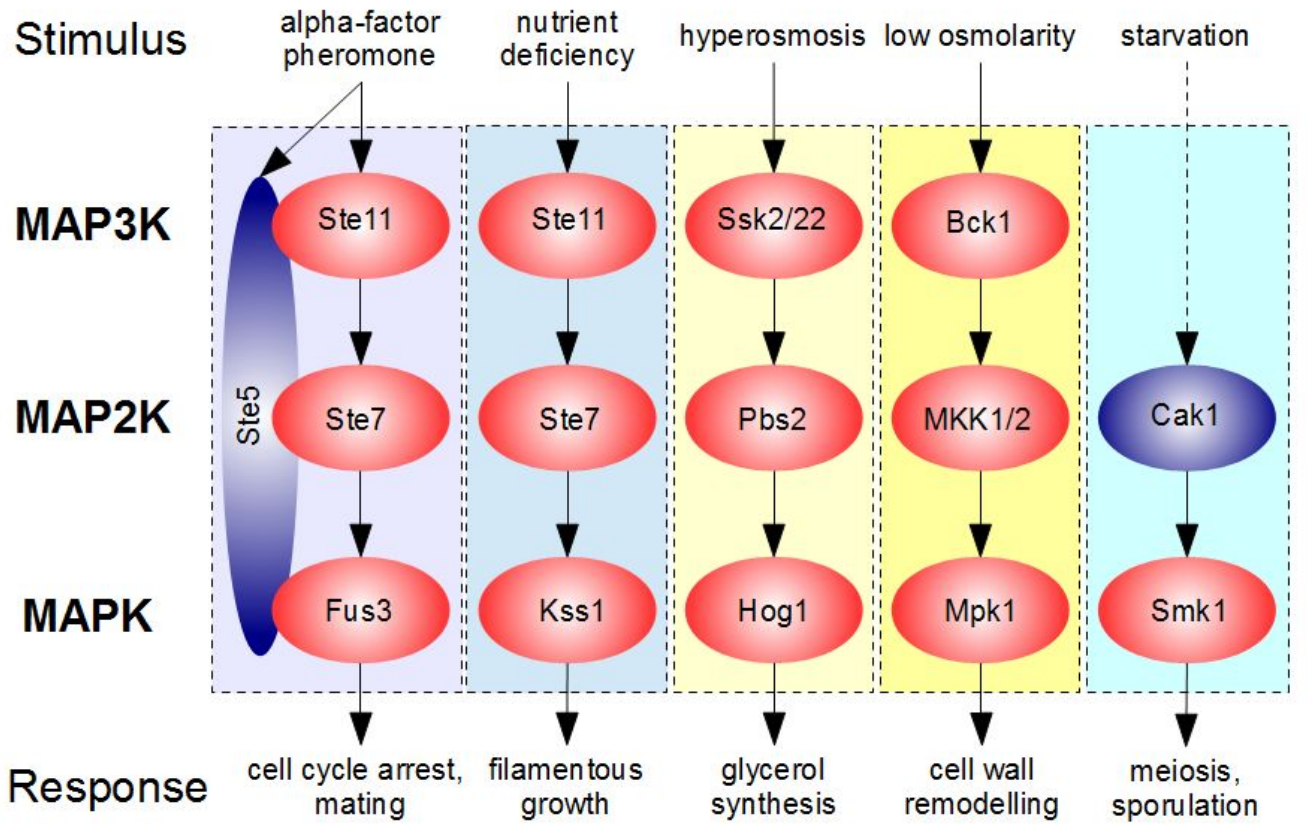


Figure 3: Brief overview of S. cerevisiae MAPK pathways (MAP3Ks, MAP2Ks and MAPKs are shown in red, while miscellaneous proteins are coloured in blue). [Modified from Wikipedia.]

Genuine stress-activated MAPKs are also found in yeast: The **Hog1** pathway, in particular is responsible for adaptation to **high osmolarity**.[64] The exact mechanism by which proteins (Sln1 and its partners) are sensing the hyperosmotic conditions upstream of the pathway are not well understood. Nevertheless, the histidine kinase Sln1, the intermediate protein Ypd1 and the final substrate Ssk1 form a two-component phosphorelay system, controlling the associated MAPK pathway.[65] These downstream components are also reasonably well-known: They include the MAP3Ks Ssk2 and Ssk22 (Ste11 is also involved), the middle tier MAP2K Pbs2 and the MAPK component Hog1. Activation of this pathway is required for glycerol production, a non-ionic osmolyte counteracting the adverse extracellular conditions. Different from other yeast MAPKs, Hog1 has a number of dedicated

substrates, including the MAPKAPK-homologous kinases RCK1, RCK2 or the transcription factor Smp1. In *Schisosaccharomyces pombe*, the evolutionarily related Sty1 pathway responds not only to hyperosmotic conditions, but to a much wider range of stress stimuli, including oxidative stress and heat shock.[66] Sty1 is activated by the MAP2K Wis1, which is in turn activated by the MAP3Ks Wis4 and Win1. These kinases also have well-established mammalian relatives: All p38 and JNK kinases are orthologs of Hog1 or Sty1.[67] Similarly to their fungal counterparts, these mammalian kinases also have a well-conserved function in regulating the cellular responses to diverse abiotic stress stimuli. By the same virtue, the MAP3Ks Ssk2 and Ssk22 are structurally similar to and thus probably orthologs of human MEKK4, while Pbs2 directly corresponds to mammalian MKK3, MKK4, MKK6 and MKK7.

There are a number of other conserved MAPK pathways in fungi, but some of them have rather different functions than their mammalian counterparts. The same is true to the yeast kinase **Mpk1** (also known as Slt2), member of the so-called **cell wall integrity pathway**. It is activated by numerous stimuli that disrupt the cytoskeleton or the cell wall itself (e.g. chitinase treatment).[68] The small G-protein Rho1 plays a key role in activating this pathway, that consists of dedicated MAP3K (Bck1), MAP2K (Mkk1 and Mkk2) and MAPK (Mpk1) components. This pathway is absolutely required by yeast cells yeast to maintain cellular integrity under hypotonic conditions. In fission yeast (*S. pombe*), the corresponding proteins are called Mkh1 (MAP3K), Pek1 (MAP2K) and Pmk1 (MAPK), but the pathway is generally thought to be similar to that of *S. cerevisiae*. One of the key downstream substrates of this pathway is the transcription factor Rlm1 (homologous to mammalian MEF2 factors). Although its members are structurally somewhat different from the mammalian ERK5 pathway (no PB1 domains are found on these yeast kinases), comparison of sequences suggest that Mpk1 could possibly be homologous to mammalian ERK5 (involved in cardiovascular regulation), and so do functional complementation experiments (human ERK5 can partially rescue Mpk1 null mutants).[69] Like many other eukaryotes, fungal genomes also encode **atypical MAPKs**. In yeast, the atypical MAPK **Smk1** is required for **sporulation**. This appears to be activated not by a canonical three-tiered pathway so typical of classical MAPKs, but by CAK1, a member of the cyclin-dependent protein kinase (CDK) family.[70] The sporulation pathway controls meiosis in yeast as well as the synthesis and assembly of the spore wall. However, compared to other MAPK pathways, our knowledge of the control of Smk1 activity or its downstream substrates is rather limited.

Most **unicellular eukaryotes** do possess functional MAPK signalling systems, though they are rarely characterized by function. In the collective amoeba *Dictyostelium discoideum*, two MAPKs were described: The ddERK1 protein appears to be a classical MAPK enzyme, regulated by a dedicated upstream activator kinase (ddMEK1) as expected.[71] On the other hand, ddERK2 is an atypical MAPK, most similar to the human ERK7 protein. Both *Dictyostelium* MAPKs are required for proper mold development, but ddERK2 is critical for spore generation, while ddERK1 is not.[16] The genome of the parasitic flagellate *Giardia lamblia* was also found to encode two different MAPK genes (unfortunately, also named glERK1 and glERK2), with different roles and subcellular localizations.[72] One of them is a classical MAPK, while the other one is clearly atypical (again, resembling the mammalian ERK7 protein). The apicomplexan parazite *Plasmodium falciparum* (responsible for malaria) also contains a pair of MAPKs, one typical (Pfmap-1) and another atypical one (Pfmap-2), the latter being essential for its differentiation.[73] These observations suggest that probably all classical MAPK pathways descended (by duplication) from a single ancestral, three-tiered pathway. Although we know very little of their substrates and docking sites, diversification of these pathways clearly predated multicellularity. It is also likely that some atypical MAPKs are just as ancient as classical ones are, hinting at a profound structural and functional divergence from the classical MAPKs.[67,74]

Higher plants also contain a large number of different MAPK proteins, assembling diverse pathways. The genome of *Arabidopsis thaliana* encodes no less than 20 (!) different MAPK genes - far more than what most fungal or metazoan genomes have; and MAPKs are influencing almost every aspect of plant life. Though plant pathways do not directly correspond to any particular human pathway, they clearly evolved from the same ancestral eukaryotic MAPK system. Plant MAPKs are usually classified into group A, B (classical MAPKs), C (partly atypical) and D (fully atypical) MAPKs. While the atypical group D MAPKs may lack upstream activator kinases altogether, group A and B MAPKs show the same three-tiered pathway architecture as human MAPKs do.[75] The best-known *Arabidopsis* pathways are the ones mediated by either **MPK3** and **MPK6** (group A MAPKs). These kinases mediate response to a myriad of different stressors, such as osmotic stress, heat shock, cold stress or infection.[76] The second best-known pathway is the one mediated by **MPK4** (group B MAPK), involved in stress responses, jasmonate signalling and regulating growth (knockout mutants of this kinase show dwarfism).[17,77] Similarly to yeast or human proteins, classical plant MAPKs also appear to utilize docking motifs to interact with their upstream activators (MAP2Ks) and their substrates (such as WRKY transcription factors).[78]

Direct substrate recognition in the CMGC kinase group

The superfamily of canonical (or "typical") protein kinases is exceptionally rich and diverse in eukaryotes. Its members are found on all major domains of life, including prokaryotes and it encompasses no less than 450 members in the human proteome.
Mitogen-activated protein kinases belong to a major branch of canonical protein kinases termed **CMGC group** - an acronym generated from its best known members: Cyclin-dependent kinases (CDKs), Mitogen-activated kinases (MAPKs), Glycogen synthase kinase 3 (GSK3) and CDK-like kinases (CLKs). Genes coding for CMGC kinases are found in every known eukaryotic genome. The family not only includes Mitogen-activated protein kinases (MAPKs) and their closest relatives, the Cyclin-dependent kinases (CDKs), but also Glycogen-synthase-kinase 3 (GSK3), Casein kinase 2 (CK2), Homeodomain-interacting protein kinases (HIPKs), Dual tyrosine regulated kinases (DYRKs), CDK-like kinases (CLKs), Serine/arginine rich protein kinases (SRPKs) and the cilium-regulating Ros cross-hybridizing kinases (RCKs: MAK, MOK & ICK). Normally, all these proteins are strictly serine/threonine kinases. They phosphorylate tyrosine side chains only under exceptional conditions (the latter is only thought to be possible when the kinase is acting intramolecularly).[79]

CMGC kinases are united by a number of features, leaving little doubt about their common origin.[80] Being a **proline-directed kinase** is almost synonymous with membership in the CMGC group. This refers to their very characteristic substrate specificity: strongly preferring a poline amino acid immediately after the serine/threonine amino acid to be phosphorylated. The requirement for a proline stems from structural features unique to the CMGC family. One side of the catalytic pocket is defined by an unusually positioned arginine, only allowing access to the catalytic site if the substrate polypeptide chain makes a sharp turn at its C-terminal end.[81] As a consequence, CMGC kinases strongly disfavour an alpha-helical arrangement of substrates that is so characteristic of many other major kinase families (like the relatives of Protein kinase A, Protein kinase C or Calmodulin-associated kinases).[82] Although persuasive, the previous structural argument still does not completely explain the nearly exclusive requirement for a proline (versus Gly, Ala, etc.) among MAPKs and some of their relatives. However, the structure of substrate-bound DYRK kinase betrays that the proline from the substrate is positioned planarly onto a phospho-tyrosine side chain from the activation loop of the kinase.[83] The latter hydrophobic interaction can readily explain why MAPKs (that share the same tyrosine amino acid, phosphorylated in the active state of the kinase) are nearly completely restricted to

Ser-Pro or Thr-Pro sites. MAPKs are much more strict in this regard than GSK3s that can readily accept other amino acids with small side chains (e.g. A, S, T or G), yet still prefer proline at the p+1 position. Apart from the proline, there are very few additional constraints. Some CMGC kinases may also have specific preferences for additional amino acids, being either positively (as in the case of CDKs: [S,T]-P-x-[R,K] and with DYRKs: R-x{1,2}-[S,T]-P) or negatively charged, including pre-phosphorylated amino acids (as for CK2: [S,T]-x{2,3}-[E,D,pS,pT] or GSK3: [S,T]-P-x{2}-[pS,pT]).[84,85] In rare cases, large, bulky amino acids may also be allowed in the place of the p+1 Pro (as with SRPK).[80] Still, these preferences are often far less strict than those incurred by kinases that prefer helically-arranged substrates. The result is a loose and partly overlapping consensus for substrates across the entire CMGC family.

Anatomy and regulation of classical MAPKs

Most mitogen-activated kinases consist of little more than a **kinase domain**. This is true to almost all classical human MAPKs, save ERK5 that has a long, intrinsically disordered extension. Similar extensions are more commonly encountered on atypical MAPKs, such as ERK3, ERK4 or ERK7. But even these extensions lack detectable long-range folding tendencies, thus we can safely assume that MAPKs are characteristically single-domain proteins. The kinase domains of MAPKs are fairly well-studied. As usual for canonical protein kinases, the catalytic pocket is formed in the deep cleft between the N- and C-terminal lobes. The catalytic aspartate is universally conserved, as are residues responsible for ATP coordination. The activation loop - which borders the catalytic site from below - is partially disordered and mobile in the basal state of the kinase, denying entry of protein substrates.

Two **phosphorylation** events are necessary for catalytic activity: both a threonine and a tyrosine side chain must be phosphorylated for the correct positioning of the **activation loop**. The dual-phosphorylated ERK2 structure shows that the several arginine and lysine side chains lining the catalytic region have an important role in coordinating the phosphorylated side chains.[86] These modifications are usually provided by an entirely different type of kinase (MAP2Ks or similar STE-family kinases), recruited to the MAPKs through their dedicated docking motifs. Kinase activating kinases are always special: they must be able to coordinate intimately (in an almost "jigsaw-like" manner) with the substrate kinase in order to reach these deeply positioned target sites.[87] Some MAPKs

can also autophosphorylate: ERK1 and ERK2 have fairly long activation loops, and ERK2 has also been shown to be able to phosphorylate its own activation loop through an intramolecular mechanism. However, the fully folded ERK2 kinase domain locks down the activation loop to prevent autoactivation.[88] Therefore autophosphorylation can only occur at an immature, incompletely folded state (or in appropriate mutants), similarly to DYRK kinases.[79] But the physiological significance of ERK2 autophosphorylation is questionable at best, since the cellular environment contains several phosphatases inactivating MAPKs. Similarly, certain forms of JNK2 have also been shown to be able to autophosphorylate: in this case, the mechanism is different, and appears to be intermolecular.[89] In other systems, autophosphorylation of p38α has also been detected, but the molecular details are unclear, as is the significance of these findings.[90] However, the autophosphorylation tendencies seen in some atypical MAPKs may have true physiological importance, as some of these kinases have no dedicated MAP2Ks to activate them otherwise.

Phosphorylation of MAPKs is a fully reversible process. Several different **phosphatases** are involved in the dephosphorylation and hence inactivation of mammalian MAPKs. These include Tyr phospahatases (e.g. HePTP, acting on ERK2), Ser/Thr phosphatases as well as dual-specificity phosphatases (DUSPs), capable of removing both phosphate groups from the activation loop of most MAPKs.[91,92] There is even a dedicated group of dual-specificity phosphatases, with the sole purpose of inactivating MAPKs: Called "MAP kinase phosphatases" or MKPs, they are critically important in keeping the MAP kinases inactive in the lack of external signals. These enzymes frequently harbour a separate, regulatory domain (the "rhodanese" domain) to recruit their MAPK substrates. Others lack external recruitment elements, relying on their catalytic domains to fit onto the activation loops of targeted MAPKs. Other phosphatases (like HePTP) also evolved recruitment elements binding to the same site of MAPKs, independently from MKPs.

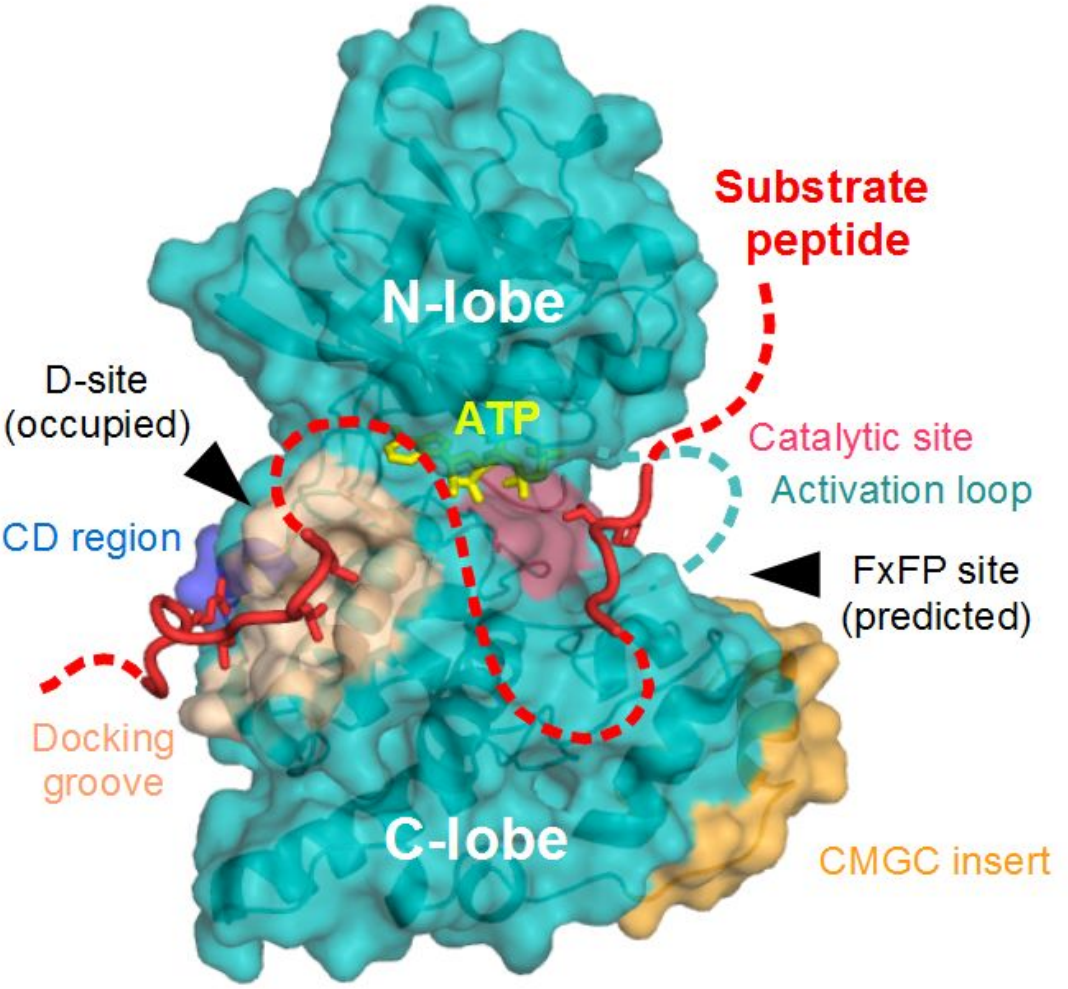


Figure 4: Structure of a MAPK during the act of substrate phosphorylation. The figure shows JNK1 bound to the docking motif of NFAT4, with the target site modelled after the DYRK1A-substrate complex. The structure of FxFP-site associating motifs is unknown, and therefore not shown here.

The C-terminal lobe of MAPKs also have a number of unusual feaures. First, the C-terminal lobe has a conserved insertion into it that protrudes from the kinase domain immediately below the activation loop. Originally called "MAPK insert", this segment is not fully unique to the mitogen-activated protein kinase family. These alpha helices are more properly termed as "**CMGC insert**", as the same structural element is found in all CMGC group members. Although relatively loosely folded and typically having high B-factors in X-ray structures, the latter site has enormous importance for MAPKs: This provides the lining of the **minor docking site** (the so-called FxFP site).[93] The C-terminal

lobe is also longer than the core kinase fold: After the last core helix, the C-terminal segment runs upward, and directly contacts the N-terminal lobe with a long alpha-helix. The extreme C-terminus of MAPKs is hence radically different from the related cyclin-dependent kinases or glycogen-synthase-kinase 3. In its relatives, the ultimate segment of the protein binds onto the C-terminal kinase domain, right under a prominent loop protruding from the C-lobe. This arrangement (that resembles a "sleeping swan", with its head tucked under its wings) no longer applies to MAPKs: the space under the β7-β8 loop is left empty, leaving a widely open hydrophobic groove behind. But the elongated C-terminus also generates a short mini-helix nearby, which is often strongly charged (the so called CD or complementary docking site). The CD region is adding on to the adjacent hydrophobic groove, giving rise to the **major docking site** (called the D-site) - fully unique to MAPKs.[94] In turn, the long **C-terminal helix** attaches to the N-lobe of the kinase domain. This attachment (that happens partly on the same surface where CDKs recruit their cyclin subunits) has two important consequences. First, MAPKs do not require cyclin subunits to stabilize their structure, unlike cyclin-dependent kinases. Second, they cannot rely on regulatory subunits for substrate recruitment or regulation of activity. The sole source of allostery required for MAPK activation must stem from its activation loop phosphorylation.[95]

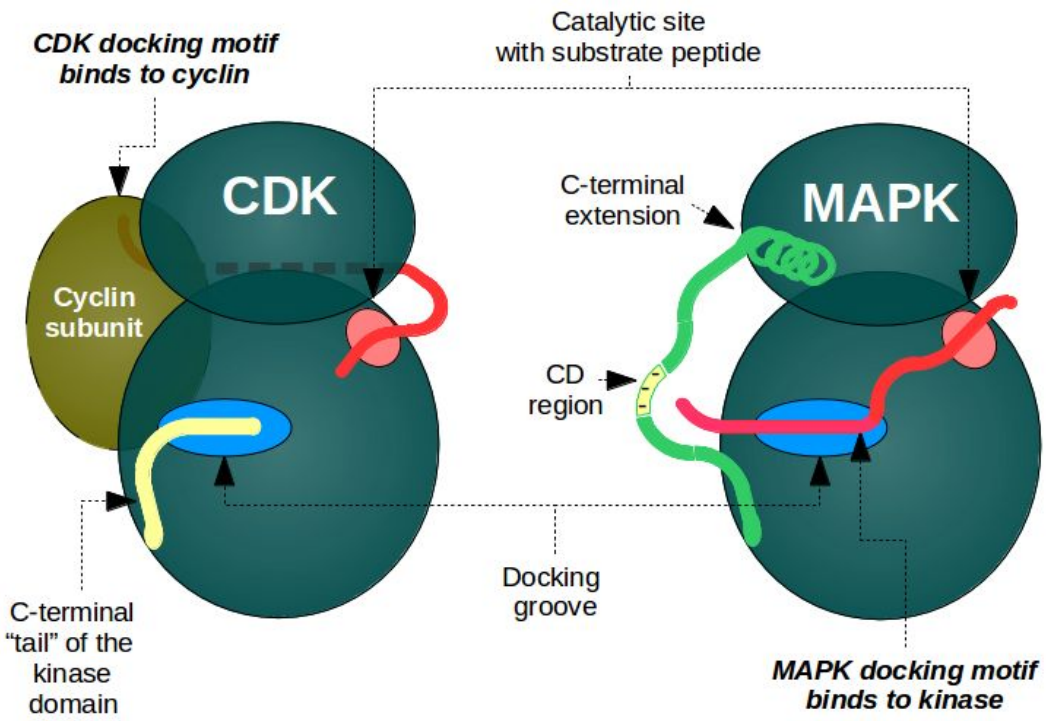


Figure 5: Schematic comparison of docking sites and substrate recognition by a cyclin-dependent kinase (left) and a mitogen-activated protein kinase (right)

Docking phenomena among protein kinases

When speaking about enzymes modifying other macromolecules, the term "**docking**" refers to substrate recognition outside the catalytic region. The extra docking surfaces responsible for precise substrate recognition are typically discontiguous with the immediate neighbourhood of the catalytic site. More often than not, docking phenomena occur in a highly modular fashion, with either dedicated interactor domains or short linear nucleic acid / protein motifs being recruited. The enzymes themselves may also utilize separate domains to bind motifs or domains found within target substrates. In other cases, recruitment of the substrate occurs on the catalytic domain itself, but on separate, non-catalytic binding sites.[96]

Docking phenomena are almost universal in enzyme systems responsible for ubiquitinylation of proteins or modification of chromatin (histone methylation, acetylation, etc). They are also extremely common among protein phosphatases. However, docking is observed less frequently among proteases, nucleases or protein kinases. The main purpose of docking is to complement the catalytic site in substrate recruitment. However, the recruitment of additional domains or linear motifs far away from the target site not only serves to strengthen the enzyme-substrate interaction: it also provides substrate specificity. Thus it is no wonder that docking phenomena are more common with enzymes possessing relatively loose primary substrate specificity (having a promiscuous catalytic site).

Protein kinases that utilize docking frequently have one or more dedicated domains to recognize the corresponding linear motifs in their substrates. These can also conveniently function as autoinhibitory domains. In the **Src family** of tyrosine kinases, the kinase domain is autoinhibited by a cooperation of the SH2 and SH3 domains. On the other hand, Src kinases preferentially recognize substrates that have proline-rich motifs acting as ligands for its SH3 domain. Pre-phosphorylation of the substrate on appropriate tyrosine residues can potentiate the activity of Src kinases even further.[97] The compatibility of the SH2 domains with motifs phosphorylated by the Src kinases themselves ensures that these enzymes can effectively act as amplifiers in Tyr-phosphorylation-based pathways (such as in cell-cell adhesion).[98] Similary to the previous example in principle but not in structure, the cell division controlling **Polo kinases** (PLKs) also possess a dual-purpose regulatory domain.[99] These Polo-box domains inhibit the activity of the kinase domain by direct binding, however they are also critically important for recognition of substrates by binding to a docking motif. (The cited motifs must be pre-phosphorylated by CDKs in order to become a Polo-box ligand - causing PLKs to act downstream of

cyclin-dependent kinases.) Unrelated to the previous ones, the **OSR/SPAK** kinases also utilize docking to connect with their upstream activators as well as substrates. OSR1 and SPAK are two closely related human kinases that act as effectors of the WNK (with-no-lysin kinase) pathway, critical in maintaining osmotic balance and regulating renal salt secretion/reabsorption. Their unique C-terminal domain directly binds to short linear motifs located on the C-terminus of WNKs; but the same motifs are also found in many OSR/SPAK substrates, including several K^+/Cl^- and Na^+/Cl^- cotransporters.[100]

In other kinases, docking predominanly occurs on the kinase domain itself. Within the **CMGC kinase group**, the loose substrate specificity (in many cases, only the sequence SP/TP is strictly required) has resulted in a rich world of docking motifs and recruitment sites. However, each kinase subfamily seemingly invented docking independently from others. In the case of GSK3 kinases, many substrates can rely on priming phosphorylations for direct recruitment to the catalytic site and thus not all utilize docking. Those proteins that do (most notably, FRAT and Axin1), use a highly helical motif tightly associating with a binding site located on the CMGC insert of GSK3.[101,102] This docking mechanism is however, very conserved, and found not only in multicellular animals but also in higher plants.[103] Although the SRPK kinases were also found to utilize a site near their CMGC insert, it is located on the opposite side and the exact geometry of SRPK docking motifs is strikingly different from GSK3: They bind to motifs that have several Arg or Lys amino acids alternating with others, e.g. RERSPTR, found in the SR2 protein.[104]

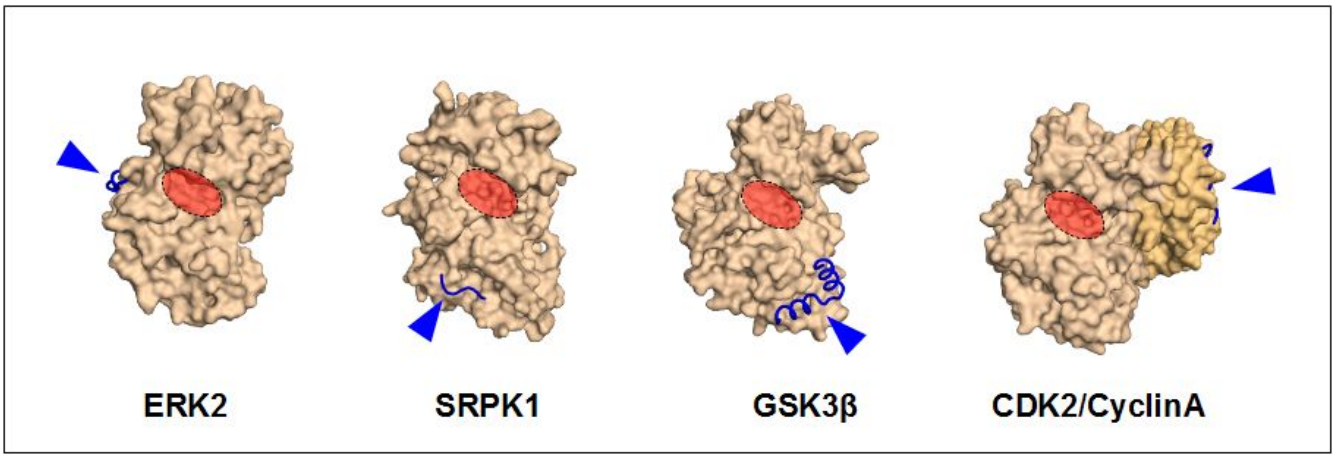


Figure 6: Many different docking sites mediate substrate recruitment among CMGC kinases.All kinase domains were drawn in the same orientation, with the ATP-binding pocket shown in red.Blue arrows indicate the position of the docking motifs on the kinase domains (wheat) or extra subunits (orange)

In the case of cyclin-dependent kinases (CDKs), docking is not mediated by the kinase domain. Instead, it is the cyclin subunits that took over the role of substrate recruitment by auxiliary interactions. In the case of the best known cyclin, cyclin A, the docking groove can accommodate motifs with the core sequence **RxL** [In truth, the better known cyclin A docking motifs fit one of the two possible patterns, either $(\theta/\varphi\text{-}\theta)_{opt}$-**θ**-x-$\varphi_1$-x-$\varphi_2$ or $(\theta/\varphi\text{-}\theta)_{opt}$-**θ**-x-$\varphi_1$-$\varphi_2$-Gly. Here **θ** stands for Arg/Lys and **φ** denotes hydrophobic amino acids, φ_2 often larger (typically Phe) than φ_1 (most commonly Leu), and the amino acids under the *"opt"* tag are optional, generating additional contacts].[105] Due to the cyclin being positioned beside the kinase domain, the cyclin docking motifs are usually located *downstream* to the targeted SP/TP phosphorylation site (the perfect consensus, [ST]Px[RK] is not always satisfied) by approximately 10-25 amino acids. However, docking phenomena vary widely within the CDK family, and there are many cyclin subunits that completely lack the binding site identified on cyclin A. The details of substrate recognition for CDKs using regulatory subunits different from the better-known cyclins is currently shrouded in mystery (as is the case for CDK5). But even for kinases like CDK2, the combinatorial association with multiple cyclins opens up the path to recognize different substrates with each different cyclin subunit (e.g. the partners recognized by cyclin A and cyclin B do differ from each other, and this is presumed to have a major impact on cell cycle regulation). [106]

Mitogen-activated protein kinases are the closest extant relatives of CDKs. However, MAPKs have no cyclin subunits: Instead they use their kinase domains extensively for substrate recognition. Two major docking sites were characterized up to date. One of them (the "minor" docking site, or FxFP site) lies at the junction of CMGC insert and the activation loop. The other (the "major" docking site or D-site) is located almost on the opposite side of the kinase domain, in the place that is normally filled out (in most CMGC kinases) by the extreme C-terminus of the protein. The two possible docking sites of MAPKs result in two, very different docking motifs. One of them is short, incorporating at least two, large hydrophobic amino acids (known as "FxFP" motif, as this is the near-optimal consensus of the "minor" docking site, determined for ERK2). The other one accepts longer linear motifs with positively charged amino acids on one end and a series of alternating hydrophobic residues on the other; with variable arrangements (these "D-motifs" - superficially similar to the cyclin docking motifs - are widely used by many MAPK partners to provide access to the "major" docking site of MAPKs). Due to the

topology of MAPK docking sites, the D-motifs are usually located *upstream* of the targeted SP/TP motif, and relatively loosely coupled to it (often separated by ~10-50 amino acids), while the FxFP motifs tend to be located *downstream*, with much more tight coupling (~5-10 amino acids below the target site). MAPK substrates are free to use either docking motifs or both (although the latter is rare) for efficient phosphorylation, as the D-site and FxFP site act in a combinatorical and synergistic manner.[107,108] Enzymes and other proteins regulating MAPK activity often utilize the D-site, with motifs surprisingly similar to those of the substrates, despite the fact that in this case the enzyme-substrate relations are swapped (the MAPK becomes the substrate). This leads to widespread competition for access to MAPKs through their major docking site (see figure).

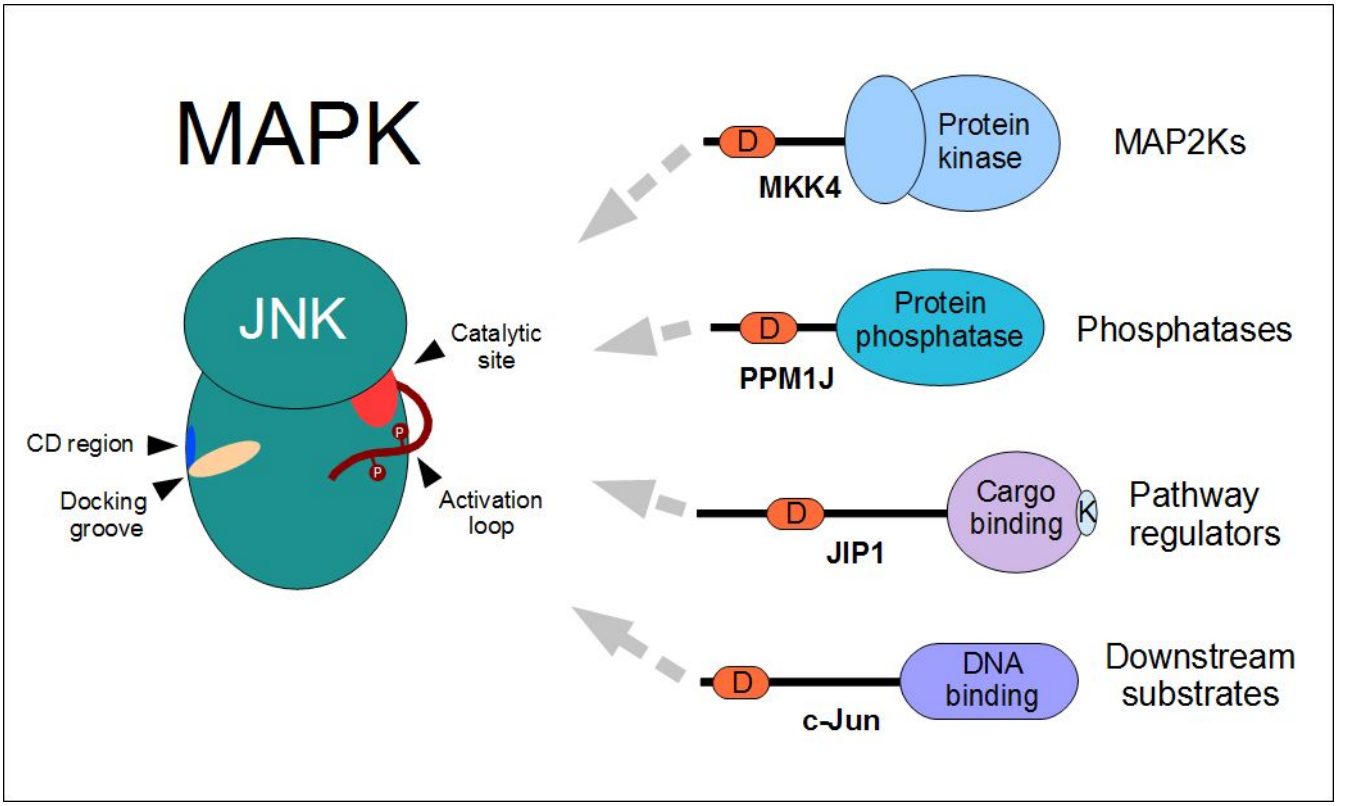


Figure 7: The major docking site of MAPKs is utilized by not only by substrates, but also by a rich variety of kinase regulators, including MAP2K kinases delivering activation loop phosphorylation and MAPK phosphatases dephosphorylating and inactivating these enzymes. Therefore all these proteins are subject to competition with each others.

Structural diversity in MAPK docking

MAPK-interacting D-motifs were first identified in the late 1990s, when researchers noticed the sequence similarities between the N-termini of MAP2Ks with recruitment regions of substrates. As several human MAPKs could be crystallized relatively easily, many X-ray structures (incorporating diverse ligands) were determined up to the present. Surprisingly, almost every attempt to co-crystallize a MAPK with a docking motif resulted in a complex with novel geometry. Thus while the similarities were notable on the sequence level, the structures were simply too different from each other to allow drawing universal conclusions. Although is true that some X-ray structures are the result of extensive manipulation and likely not representative of the true biological assembly (as with the MKP3-ERK2 "complex", ERK2 crystallized with a peptide torn out of a large, well-folded rhodanese domain)[109]; the majority does appear reliable in a biological sense. The structures of several yeast (*Saccharomyces cerevisiae*) motifs have also been available for some time, owing to earlier studies of Attila Reményi.[110] But these motifs, appeared unique, too. Only in the past few years did we obtain enough structures to begin to detect consistent structural similarities between motifs from different proteins. During my work on our previous research project, my colleagues Imre Törő and Gergő Gógl successfully determined the structure of several novel (more than 4) MAPK+D-motif complexes.[31] This was crucial in order to achieve the breakthrough in a structural classification of D-motifs.

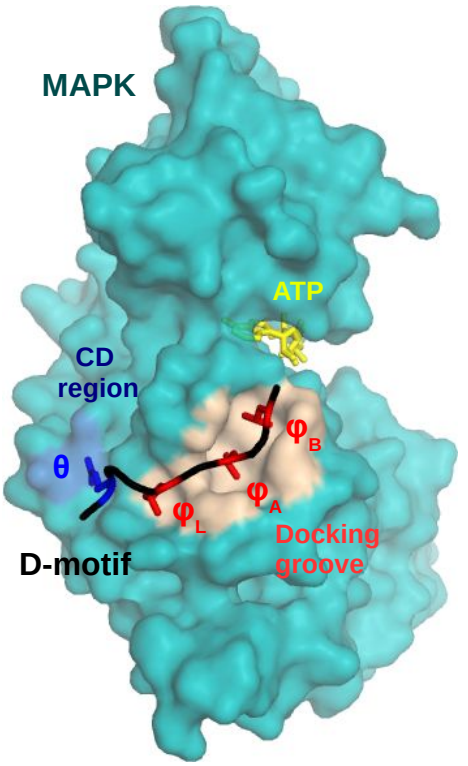


Figure 8: Generic structure of D-motifs, shown by the example of the NFAT4-JNK1 complex. The motifs typically bind a bipartite manner, with charged (blue) residues contacting the CD region and hydrophobic (red) residues binding to the docking groove. The D-site is not located directly at the catalytic site (indicated by the ATP analogue AMPPMP bound to JNK)

Perhaps one of the most striking revelation was that the D-site of MAPKs do not even bind all peptides at the same **direction**. We could show that there exist a minor group of docking motifs that bind in a completely reverse orientation to the same site. These elements, termed reverse D-motifs (RevDs) run in a C-to N-terminal orientation, compared to the ordinary N to C orientation of ordinary D-motifs, from the CD region to the end of the docking groove. Consequently, in RevD motifs, the N-terminus contains the alternating hydrophobic residues, and it is their C-terminus that is positively charged. The number of known reverse motifs is much smaller than canonical D-motifs; yet they encompass both relatively short (as in the C-terminus of PEA-15) as well as extensively long motifs (e.g. motifs from the RSK-MAPKAPK family of kinases).

Another key observation of ours was that most docking motifs use not two, but **thee hydrophobic amino acids** to contact the docking groove. Due to a few unreliable structures published before, previously it was believed that only two contacts ("LxL motif") are required. This has led to a high number of false alignments and assumptions in the past. In truth, all three pockets are filled with a hydrophobic residue in the absolute majority of cases, but not all are required to be leucines. The lower or "L" pocket (that was not recognized before) may accept either Leu, Ile, Val, Met or even Pro as shown by numerous different structures. The middle or "A" pocket is the most conservative: in most cases, this is filled out by a leucine residue, but in rare cases, Ile or Val are also detected. The rightmost or "B" pocket is not much of a pocket, but rather, a partly open hydrophobic surface. As a consequence, it can not only bind to Leu, Ile or Val, Met or Pro, but even Phe amino acids. This unexpected "permissiveness" of pockets explain why this triple-contact system was not described before. And there is one more reason: While the distance between amino acids contacting the A and B pockets is always 2 (i.e. A-x-B), the distance between the L and A pockets can vary between 2 or 3 (i.e. either L-x-A or L-x-x-A), depending on the actual motif studied. This already results in a variable hydrophobic consensus (either L-x-A-x-B or L-x-x-A-x-B), but the majority of variability does not arise from here: Instead, it stems from the linker region and the N-terminus.

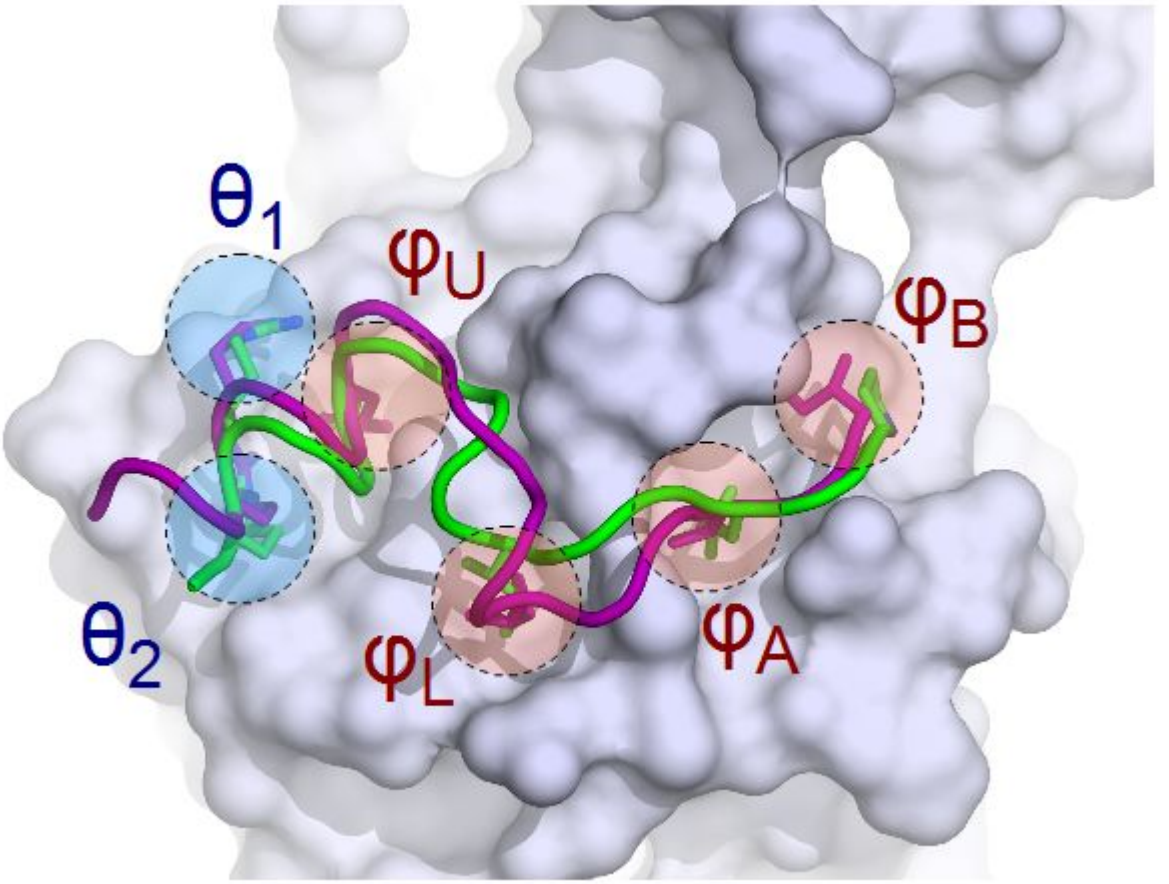


Figure 9: Despite obvious sequence and structure variations, D-motifs tend to contact the surface of MAPKs at the same contact points, including mostly charged contacts at the CD-region (θ positions) as well as hydrophobic contacts at the docking groove (φ_L, φ_A and φ_B positions). On the figure, two long RevD motifs are shown: RSK1 (green) and MAPKAPK2 (magenta), over the surface of ERK2.

The **length** of D-motifs is just as variable as their arrangements: The the docking motif of JIP1 (determined in complex with JNK1) is probably the shortest complete D-motif in existence, with only 7 amino acids required for binding (The core sequence is: RPTTLNL, with critical amino acids underlined). On the other end of the spectrum, we find the excessively long motifs from HePTP, RSK1, MNK1 or MAPKAPK2. The latter currently holds the record of being the longest D-motif with a known structure, its core consisting of no less than 17 amino acids (IKIKKIEDASNPLLLKR). Yet there are many other examples (e.g. the D-motif of *S. cerevisiae* Pbs2 kinase or the RevD motif of the *D. melanogaster* Capicua protein), that are probably even longer.[111,112]

Despite the variations in length and internal arrangements, measurements by my colleague Ágnes Szonja Garai pointed out that there are still only two fundamental groups of mammalian docking motifs, as far as **specificity** is concerned. One group of motifs interacts with JNK kinases at high affinity, while they frequently have low or no affinity towards ERK2 and p38α. The other group, which shows the ability to bind both ERK2 and p38α (and even ERK5) with biologically relevant affinities, but frequently without the capacity to interact with JNK1. Indeed it turned out that very few motifs are capable of discriminating between ERK1/2 and p38 kinases. Only the longest D-motifs and RevD motifs were capable of doing so. This suggested that the D-sites of ERK1/2 and p38s are fundamentally similar, but in some manner, very different from JNKs.On a closer observation, this appears to be caused by slight changes to the CD region in JNK kinases. Notably, a Glu side chain in the nearby ED (extended docking) region that is preserved in almost all classical MAPKs, has been changed to Lys in JNKs. This likely created an electrostatic attraction between the CD and ED regions (instead of repulsion), resulting a slight re-positioning of the CD helix. With their "collapsed" CD-region, JNKs are able to interact with motifs that no other MAPKs can: We have observed that most of the JNK-interacting motifs have fewer amino acids in-between the last hydrophobic amino acid contacting the docking groove and the first positively charged residue. As a "rule of thumb", this linker region of motifs is approximately 1 to 2 amino acid shorter in dedicated JNK interactors than motifs cognate with other MAPKs. However, their conformations are also often different. Knowing these differences, it was even possible to switch the preference of a promiscuous peptide (the D-motif of MKK4, binding to all three MAPKs) to a p38/ERK-only version, by introducing a proline into its linker region (thereby forbidding a helical conformation that its linker should adapt when binding to JNK). Similarly, we could design promiscuous peptides from existing ones or even from a scratch, by adding amino acids to positions favourable for JNK as well as into those ones required by p38α or ERK2.[31]

Surprisingly, we were able to confirm that certain motifs have **N-terminal regions** that can never be located in the X-ray structures. This did not appear to be an artefact, though: The same motifs were unable to bind whenever their charged residues (showing no detectable electron density) were removed. In addition, we only detected this phenomenon with shorter motifs interacting with either ERK2 or p38α, but there this behaviour was consistent. Upon tracing, these motifs were either partly invisible and overtly linear (as long as the main chain was traceable) or perfectly visible and of a helical conformation. Upon closer examination of the D-site of ERKs and p38s, this behaviour can be conveniently explained by the high change density of their CD region and its neighbourhood (the ED region). Since these motifs also tend to contain multiple arginines and lysines, the number of

conformational possibilities is growing so large (unless the motif is "locked" into a single conformation by becoming helical, for example) that the N-terminus can only be described by a broad "enseble" of structures, producing an uninterpretable electron density at the flexible region.

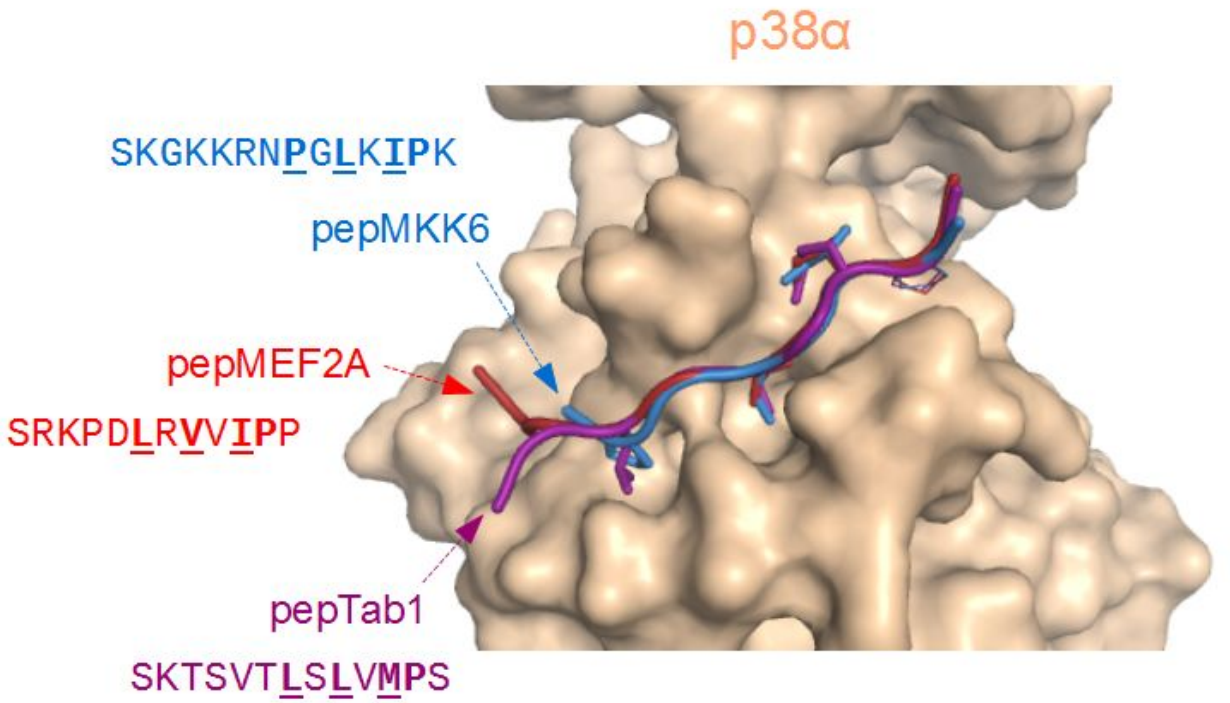


Figure 10: Three different p38α-binding motifs, showing nearly-identical C-terminal segments (the three hydrophobic contact points are shown with a stick and underlined in the sequence), but different N-terminal arrangements. The N-terminus of neither peptide is fully traceable in the X-ray structure.

Finally, some of the motif length differences can also be explained if we look at how their N-termini are arranged. The motifs of MEF2A and HePTP both bind to ERK2 with high affinity. Their orientation as well as hydrophobic residues contacting the docking groove are all similar. What is strikingly different though, is their N-terminal ends. The N-terminus of the MEF2A motif is almost completely **linear**: therefore it has the shortest distance possible for ERK2 between the two charged residues contacting the CD region and the lower pocket contact point: merely 2 amino acids. In the case of HePTP, the conformation is very different: It contacts the CD-region with a **long alpha-helix**, that is placed into the cleft between the CD sie and the docking groove (this we termed "CD groove"). An additional hydrophobic contact point is also generated by this helix: This fills out the U or "upper"

pocket on ERK2. As it is located to a considerable distance from the lower pocket, a linker of 5 amino acids is utilized in the HePTP motif to traverse the space. The two charged amino acids contacting the CD region are part of a large helix starting at the U point and winding downwards, until the end of the CD groove. A helical arrangement of the N-terminus is entirely facultative: some motifs utilize it (and these end up being substantially longer) in exchange for somewhat higher affinity and specificity, while others have N-termini running linearly, avoiding the U pocket. Due to the extensive linkers observed with these "longer" motifs, there is plenty of space for additional specificity-determining contacts: And these are the ones that lie behind the potential selectivity of longer motifs between ERK1/2 or p38s. The upper pocket also exists in p38 kinases (and occasionally used as well by longer motifs), but not on JNKs. There is no place for a helix in the "collapsed" CD region of JNKs: Therefore motifs with a long N-terminal helix select strongly against interacting with JNK.

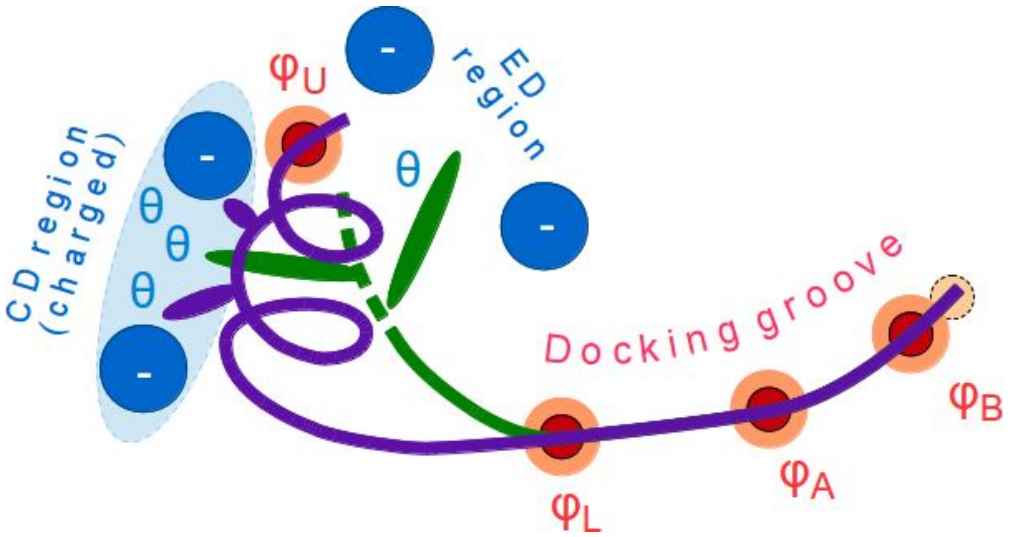


Figure 11: differences between the topologies of a linear (MEF2A, green) and a helical (HePTP, purple) motif, binding to ERK1/2 or p38, when viewed from above.

Clearly, the problem of different motifs binding to the very same site is not unique to MAPKs. Many if not most linear motif - domain interactions are expected to behave similarly. The usual approach of biologists in this case is to seek the "highest common consensus" across different motifs, without paying attention to structural differences. However, this lackluster approach tends to yield core consensus motifs that are sometimes completely wrong; But even when grossly correct, they are usually not predictive of binding capacity alone.

AIMS & OBJECTIVES

The docking phenomena in MAPKs were already studied by numerous groups before. However, the structural understanding of these interactions was meager, and the previous models could not be used for reliable predictions of MAPK-partner protein interctions. Hence our main aims were:

- To develop a **structurally consistent model** of MAPK-partner protein interactions, that could enable direct prediction of MAPK-target protein partnerships.
- To develop appropriate ***in silico*** **methods** to identify novel human MAPK partners based on their primary sequence alone.
- To develop efficient, relatively high throughput **experimental method(s)** for the identification of functional docking motifs
- To determine the **specificity profiles** of novel interactors versus ERK2, JNK1 & p38α and compare the results with the *in silico* predictions
- To study these interactions in cells, with full-length proteins in order to **validate** the results of fragment-based assays.

By the analysis of our primary hits and refined models, we sought answers for a number of intriguing biochemical problems, namely:

- To identify the likely **physiological function** of docking motifs in MAPK interactors, as well as to uncover the pathological role of MAPK interactions in certain diseases.
- To develop a generic biochemical **paradigm** of MAPK-dependent regulation of substrates. This could be used in the future to identify “phospho-switches” underlying MAPK signalling.
- To analyze the **evolutionary origins** of docking interactions and the molecular mechanisms by which they emerged.
- And finally, we also aimed to study proteins that turned out to be **outliers** from our model. This was done in order to understand the limitations of a systematic, primarily structure-based approach.

RESULTS

Development of a unified structural model for MAPK docking motifs

As it was outlined in the introduction, the structures of MAPK-associating D-motifs are remarkably varied and complex. Unfortunately, such complexity was a major hindrance to develop structurally consistent consensus motifs; however, such a model is necessary for reliable *in silico* identification of MAPK partners. The currently "most widely accepted" D-motif consensus is available at the ELM (Eukaryotic Linear Motifs) database as **DOCK_MAPK_1.**[113] Its definition as a regular expression is the following: [KR]{0,2}[KR]-x{0,2}-[KR]-x{2,4}-[ILVM]-x-[ILVF]. Although it was based on a number of validated examples as well as various complex structures, this motif is still very inconsistent. One of the key structural problems with DOCK_MAPK_1 is that it only requests for two hydrophobic amino acids, while the true complex structures almost always contain *three* hydrophobic contacts. The other problem relates to the flexible number of intervening residues. As our earlier research has shown, the number of intervening residues (between the hydrophobic and charged positions) is not arbitrary. It critically determines the identity of MAPK partners. As a rule of thumb, JNK prefers motifs with short linkers, while more prototypical MAPKs, such as ERK or p38 require at least one amino acid longer linkers. Due to these issues, predictions with the DOCK_MAPK_1 cannot be considered reliable (it also fails to detect a number of well-known positives, such as the D-motif of the HePTP phosphatase).

Despite the fact that D-motifs are indeed very variable, one observation helped us to reveal a possible solution. Early during my work, I observed that most of the known JNK-binding linear motifs obey one of the two most common models. Because both have well-established structural templates, we will henceforth refer to these as the "JIP1-type" and "NFAT4-type" motifs (see figure).
The fact that the overwhelming majority of JNK-associated D-motifs can be described by just two, well-characterized sub-consensus suggested that the same sub-classification might also be applicable to ligands of other MAPKs. In turn, a meaningful set of subclasses could improve the prediction of novel D-motifs immensely. Although a protein surface can be targeted by flexible linear motifs in a wide variety of ways, the number of tightly binding, conformationally different solutions might be very limited. The motifs could then be defined as more-or less regular variations of a few main solutions.

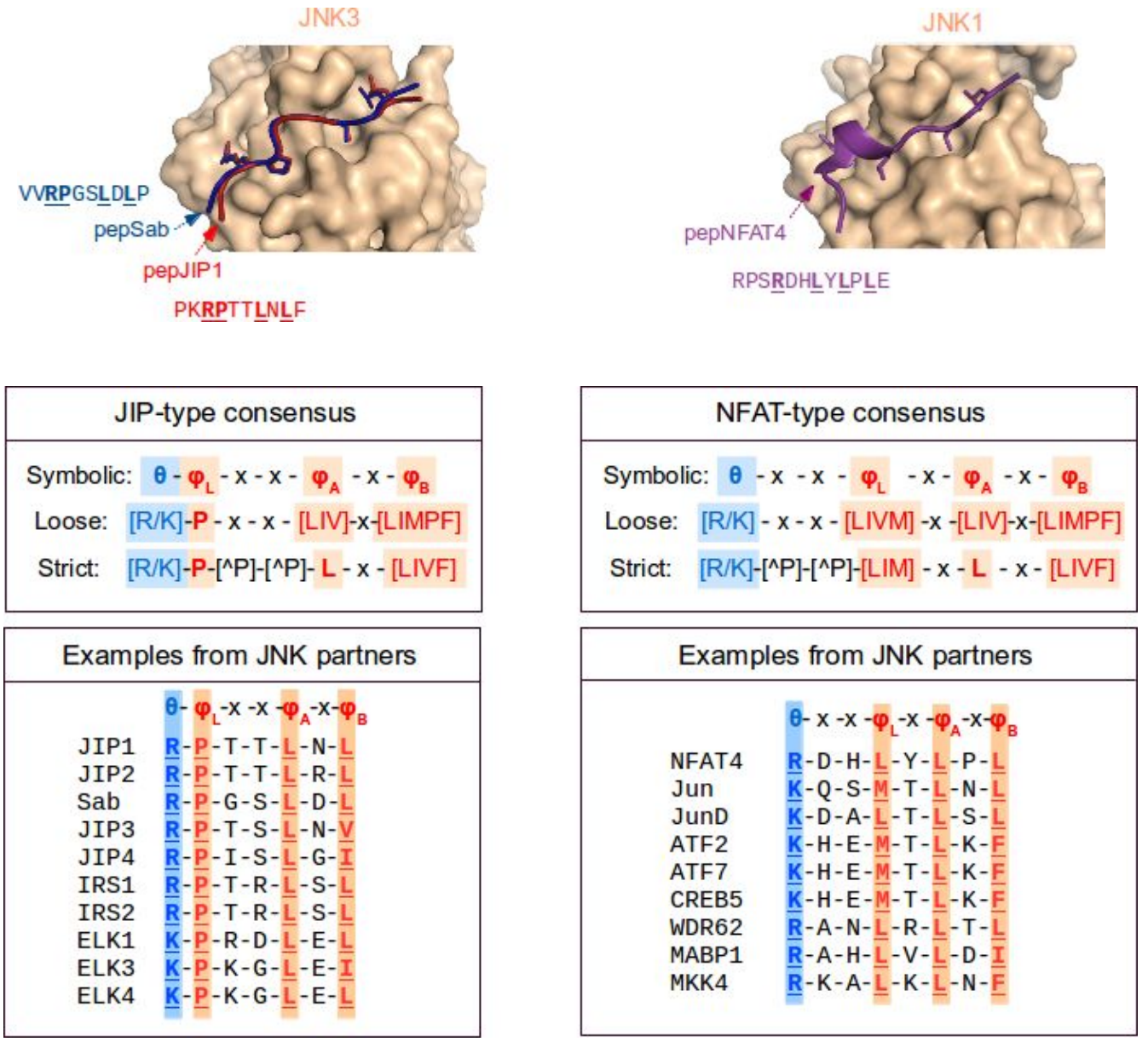


Figure 12: Sequence- and structure-based classification of previously described human D-motifs interacting with JNK. The "strict" consensus is derived from my experiments in the current study. Note that not all known JNK-docking elements are shown.

Regarding JNK only, the picture is relatively clear. Comparison of D-motif sequences suggest that most of the known JNK-interacting motifs must satisfy either the **θ**-φ_L-x-x-φ_A-x-φ_B (**JIP1-type**) or the **θ**-x-x-φ_L-x-φ_A-x-φ_B (**NFAT4-type**) consensus. Here, the sequences are displayed in a "symbolic notation" for easier reading than the usual regular expressions. The φ_L , φ_A and φ_B letters denote positions that are typically filled by hydrophobic amino acids. The lower indices L, A and B refer to lower pocket, and pocket A or B, respectively. The **θ** positions are always positively charged (Arg or Lys) while "x"

denotes any amino acid. Fortunately, a structural template is available for both classes. The JIP1-type binding is exemplified by the structures of the JIP1-JNK1, JIP1-JNK3 and Sab-JNK3 complexes (all extremely similar to each other), while the NFAT4-type binding is more-or less accurately represented by the single NFAT4-JNK1 complex.[31,114,115] The critical difference between these two binding modes lies in the positioning of the hydrophobic residues. The number of intervening residues between the φ_L and φ_A positions is *one* in the case of the NFAT4 motif but *two* with JIP1-type motifs. This, in turn, also determines the placement of the charged residues N-terminal to the hydrophobic region; In the case of NFAT4, there are two residues between θ and φ_L , forming a 3-10 helix, while in JIP1, θ and φ_L are immediately found next to each other, with no residues in-between the two. Due to such internal correlations, these two motifs cannot be described by the same consensus; they have to be handled as entirely separate types of linear motifs. This is the first point where we have to split the definition of D-motifs, despite the fact that 1-spacing and 2-spacing motifs do not necessarily have an impact on partner selectivity. The same MAPK (in this case, JNK) may bind peptides of either type. Importantly, the same "1-spacing/2-spacing" dichotomy can also be observed in motifs associating with ERK1/2 or p38s. The residue count between the φ_L and φ_A positions is *one* in the case of docking motifs found in MEF2A, MKK6 or HePTP, but *two* in the case of a motif derived from the DCC protein.[31,116–118] This duality even extends to yeast proteins: interactors of the Fus3 MAPK may have a motif with 1-spacing (Ste7, Msg5) or 2-spacing (Far1).[110] The docking surface of classical MAPKs is ancient and well-conserved across all eukaryotes. Thus it is possible for a yeast kinase to interact with D-motifs in a similar manner as its human relatives do.

Because the CD region of ERK and p38 is wider compared to that of JNKs, the N-termini of the former motifs have a much larger conformational freedom. The N-termini of these peptides are frequently not visible in the X-ray structures - even when bound to a MAPK. (Earlier experiments unanimously suggest that these residues are still critically required for binding.) Having more than one charged residue (θ) also appears to be the norm among motifs interacting with ERK1/2 or p38s. These charged positions in the motifs also show a certain evolutionary variability, with complex and often different sequence arrangements. Despite the high number of possible variations, I observed that certain arrangements were more common than others. The N-termini of MEF2A, MEF2C and MKK4 motifs show the preferential positioning of two charged residues next to each other: θ_1-θ_2-x-x-φ_L-x-φ_A-x-φ_B. (**MEF2A-type**) On the other hand, the comparison of MKK6 and MKK1 motifs hints at another possible arrangement: θ_1-x-θ_2-x-x-x-φ_L-x-φ_A-x-φ_B (**MKK6-type**), both being subtle variations of the same base model (greater MEF2A class). These were by no means the only observed patterns, but the

only ones we had a structural template for. Using these two sub-models, we could reduce the number of corresponding hits during *in silico* searches without loosing any of the well-characterized examples.

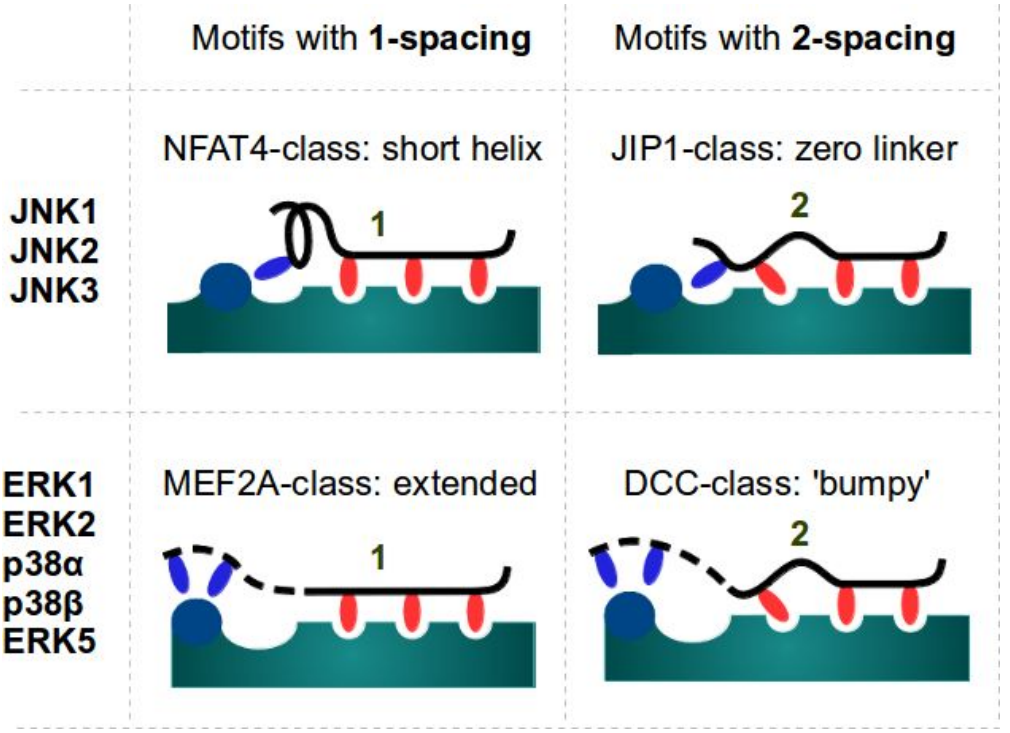


Figure 13: The free combination of hydrophobic residue spacing with linker length variations - determining the identity of interacting partners - give rise to four main docking motif classes (not considering the linear/helical dichotomy in motifs interacting with ERKs and p38s).

The structure of the HePTP-ERK2 complex serves as an important example for even longer motifs. It displays a helical structure with two charged amino acids ($\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$) and an additional hydrophobic contact ($\boldsymbol{\varphi}_U$) (where U refers to upper pocket). The **HePTP-type** docking motif (definition *sensu stricto*: $\boldsymbol{\varphi}_U$-x-x-$\boldsymbol{\theta}_1$-$\boldsymbol{\theta}_2$-x{5}-$\boldsymbol{\varphi}_L$-x-$\boldsymbol{\varphi}_A$-x-$\boldsymbol{\varphi}_B$) appears to be restricted to a single family of phosphatases, interacting with ERK1/2 or p38s.[117] The HePTP (PTPN7), STEP (PTPN5) and PTPRR phosphatases are the only known examples for this type of D-motif, and they all share a common evolutionary origin.[119] Although the HePTP model appeared to be unique, we noted that the helical N-terminus is also a property of the yeast Ste7 and Msg5 docking motifs. These non-mammalian motifs were thoroughly characterized, including structure determination by X-ray crystallography. Despite the undeniable similarity, there are still many notable differences between the motifs of Ste7 and HePTP. Apart from the different linker length, their helices are also slightly differently positioned. Interestingly, Ste7-derived peptides are known to bind human ERK2 with high affinity.[108] Therefore, we also set up a

hypothetical subclass of **Ste7-type** motifs (φ_U-x-θ_1-θ_2-x{4}-φ_L-x-φ_A-x-φ_B), hitherto unknown in humans.

There are only two mammalian examples for the ERK1/2 and p38-interacting motifs with 2-spacing between φ_L and φ_A: the **DCC**-ERK2 and MKK2-ERK2 complexes. Here, the N-termini of motifs are also flexible. Comparison of these structures with the yeast Fus3-Far1 complex and further fungal motifs with 2-spacing (Dig1, Dig2) also suggests high conformational variability, just like for motifs of the greater MEF2A class. However, a polyproline helix-like conformation between φ_L and φ_A positions appears to be a common feature. An initial DCC-type consensus motif (θ-$x_{(2-4)}$-φ_L-x-x-φ_A-x-φ_B) was thus set up to search for such motifs in the human proteome, however, this generic motif was later split up into two subtypes after some experimental testing (see figure).

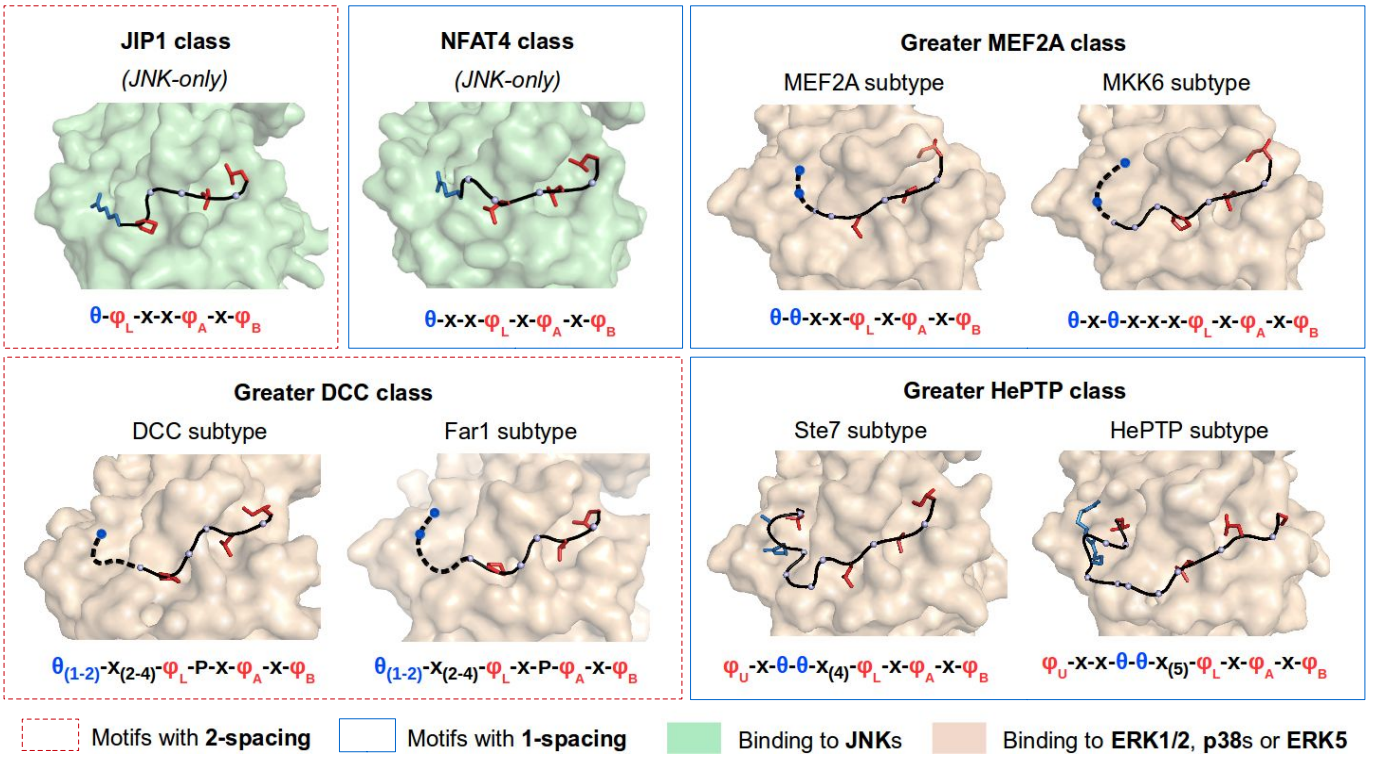


Figure 14: The structure and consensus of major D-motif classes, with their better-known subtypes and variations.This shows 5 out of the 6 theoretically possible motif arrangements (the sixth type, motifs with 2-spacing, long linker and helical end has only been observed with RevD motifs so far)

Compared to classical D-motifs, a “mirror image” like orientation for certain MAPK-binding motifs has also been described. In these “reverse docking” (**revD**) motifs, the hydrophobic stretch is located N-terminally to the charged residues. Apart from a short revD motif in PEA-15, all other mammalian examples are from a single group of MAPK interactors, the RSK/MAPKAPK family of kinases.[87,120] The low number of known reverse docking motifs, however, precluded their analysis in any systematic way.

An *in silico* pipeline to identify potential D-motifs in the human proteome

With a set of structurally consistent consensus motifs on our hand, we developed a complete in silico method for prediction of MAPK docking motifs directly from the protein sequences. First, a relatively simple motif matching algorithm was made, which - with the addition of several filters - was capable of separating meaningful motifs from the background noise. By definition, linear motifs are always found in an intrinsically disordered part of a protein. We found that the same is true to D-motifs: The overwhelming majority of known instances fall into protein segments predicted to be disordered by IUPRED.[121] Similarly encouraging results were obtained with the generic linear motif predictor ANCHOR:[122] that gave us the idea to use this base method for our first filter. This **primary search** algorithm was conceived and developed by Bálint Mészáros and Zsuzsa Dosztányi, together with much of the following filters. After the initial motif matching from UniProt (containing 20,248 human sequences), we used overlaps with ANCHOR regions as a filtering criterion. It turned out that the generic ANCHOR definition for motif-containing regions (score >0.5) could be further improved with a semi-adaptive condition, using additional IUPRED predictions. This initial filtering step was able to recover over 90% of known examples, with relatively high enrichment ratio over unfiltered motifs: We were able to reduce the number of initial hits considerably, resulting in an estimated enrichment factor over 3.8.

At the next step, we applied a series of **accessibility predictions**. Motifs that are located in disordered regions, but are found either in the extracellular space, inside the ER, Golgi apparatus, mitochondrial matrix or peroxisomal lumen are unlikely candidates for MAPK binding. As these kinases are only found in the cytoplasm or the nucleus and related, kinase-accessible compartments, we had to set up a

second filtering step to discard motifs with inappropriate localization. A secondary problem was the accidental compatibility of certain MAPK docking sequences with that of the signal peptides, which had to be filtered out as an additional step. The localization filter was realized by a combination of SignalP, Wolf Psort and Phobius algorithms.[123–125] While the ordinary cutoffs provided by SignalP were appropriate for use (after some optimalization), Wolf Psort did not provide such value. Therefore the cutoff values for Wolf Psort predictions had to be chosen "manually" (by using UniProt annotations) for different compartments. Finally, the topology was predicted by Phobius: here, we did not have to resort to additional fine-tuning of the algorithm, as it was already sufficiently reliable for our purposes.

Finally, we introduced a third set of filters. This was a knowledge-based algorithm, using known or predicted **domain** signatures (from PFAM-A) to remove any improperly predicted motif that would otherwise overlap with a known protein domain.[126] Although PFAM does provide several categories for its HMM-based predictions ("Domain", "Motif" "Family"), neither of them contain a priori information on the associated structures or lack thereof. Thus we had to re-assign each PFAM model to two categories ("Folded domain" or "Disordered segment"). PFAM-A domains are typically associated to structural units, therefore these ones were automatically regarded as "Folded domains" for most part. Some manual curation work revealed that the majority of "Motifs" are indeed intrinsically disordered, or at least predicted to be so. On the other hand, the "Family" PFAM-A category contained folded domains and disordered segments in nearly equal numbers: these had to be set apart on a case-by-case basis, by manual (literature- and folding predictor-based) curation. Once this has been done, however, the predictions were noticeably further improved. Last but not least, we also filtered out coiled-coils: to do that we employed the predictor software COILS.[127] Motif occurrences that overlapped with any coiled-coil region predicted by COILS (using the default cutoff) were removed as well. Motifs that passed all filters were regarded as possible candidates for functional D-motifs. With the total number of hits still over 10,000, the enrichment factor at this point was estimated to be over 5.

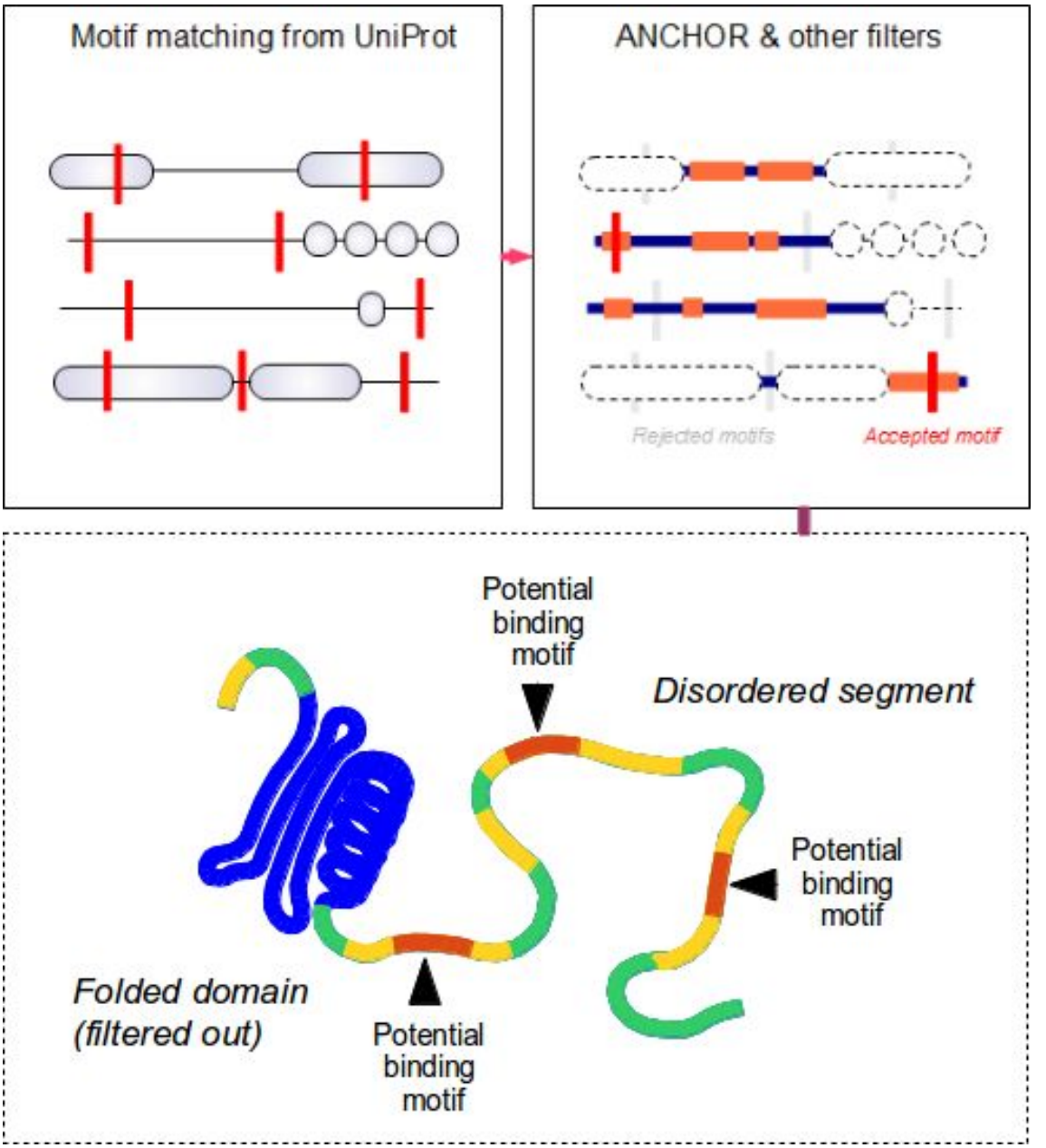


Figure 15: Schematics of the primary motif search algorithm (upper row) and graphical explanation of the ANCHOR filtering (lower row)

Estimation of binding energies using structural templates of MAPK+D-motif complexes

After the initial search step, we still recovered an overtly high number of potential D-motifs (ranging to several 1000s), making direct testing of all candidate motifs unfeasible. Thus we set up a method to estimate “goodness” of individual motifs. Unfortunately, the number of each motifs per motif class was too low to enable construction of reliable amino acid frequency (PSSM) based predictors. Thus we opted for a structural modelling based method, and used FoldX for binding energy estimation. This allowed the scoring of motifs according to their structural compatibility to the MAPK docking groove.

FoldX is a method developed to estimate the destabilizing potency of mutations introduced into a protein structure.[128] In addition, it is also widely used to estimate the binding energies of protein-peptide complexes *in silico*.[129] The main advantage of FoldX (that is also its greatest drawback) is that it assumes no change to the main chain geometry after mutation. Only the side chains are rotated in order to find the new energy minima. Therefore FoldX is expected to perform well where dealing with complexes of nearly identical geometry; performance is expected to drop where the main chain suffers changes (such can be the case when mutating Pro or Gly amino acids for example). However, this restraint in FoldX enables computation of interaction energies for a large number of potential complexes under reasonably short time. This part of the current project was executed by Olga Kalinina and Tomas Bastys, at the Max-Planck Institut für Informatik, Saarbrücken.

For each motif class or subclass, we created suitable templates from the already-published X-ray structures: these formed the basis of modelling with FoldX. Subsequently, all potential motifs were tested as “mutations” to the corresponding models and had their interaction energies estimated purely by structural modelling. From the analysis of results, it turned out that most known positives were located among the upper 50% on each lists and enriched at the top (the best performing ones being the JIP1 types). Therefore FoldX-based rankings were used as a guidance to select novel motifs for testing (we intentionally selected more motifs to test from higher ranking instances).

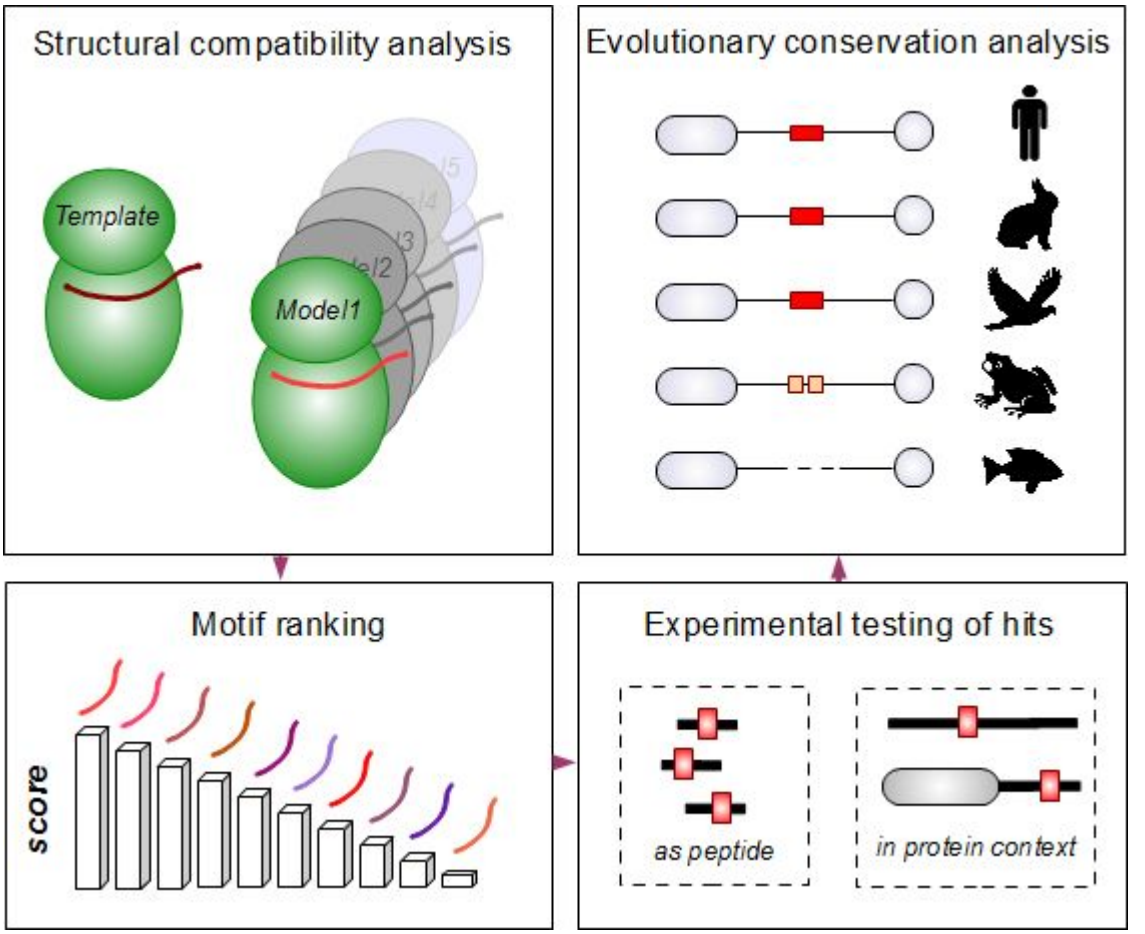
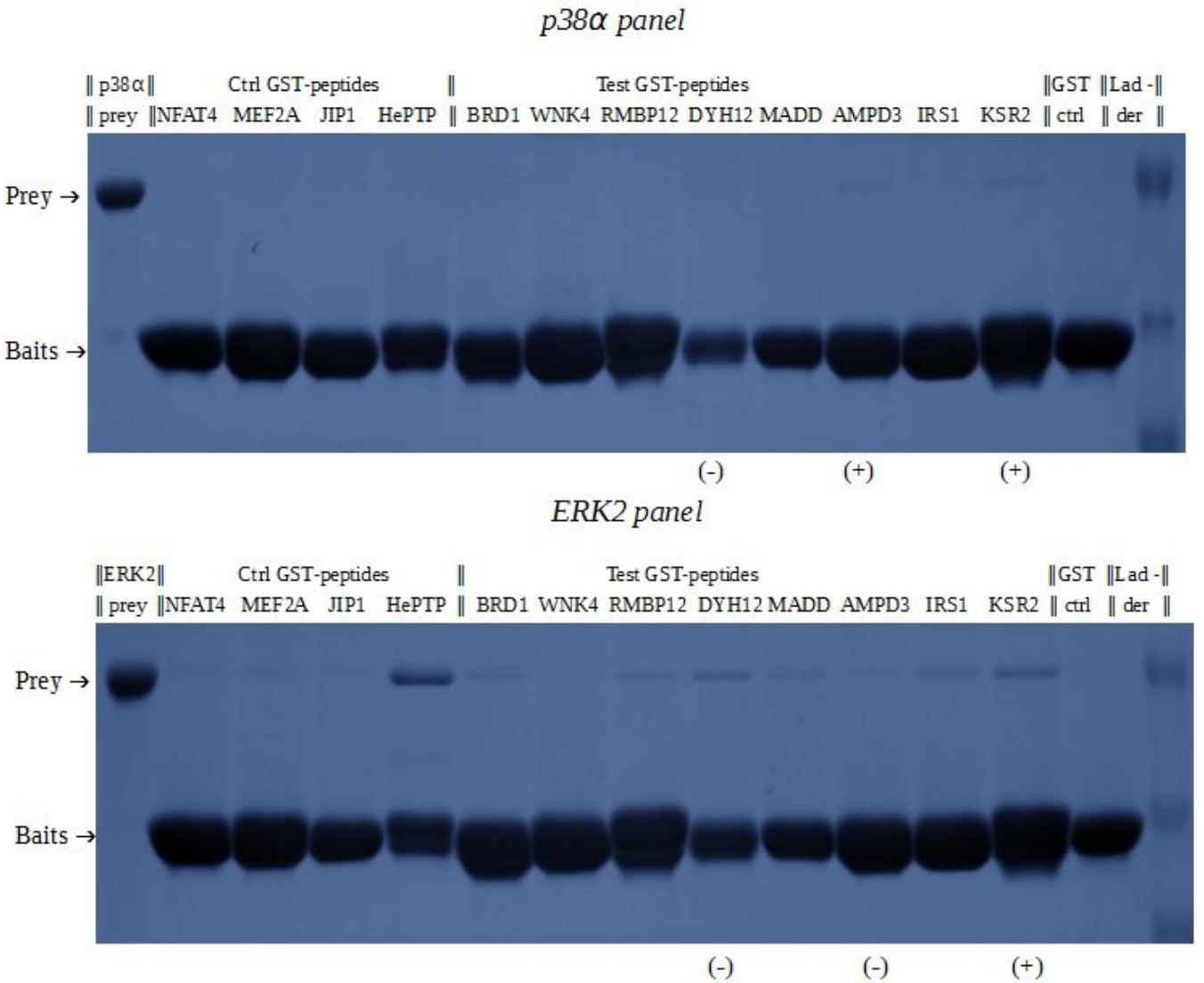


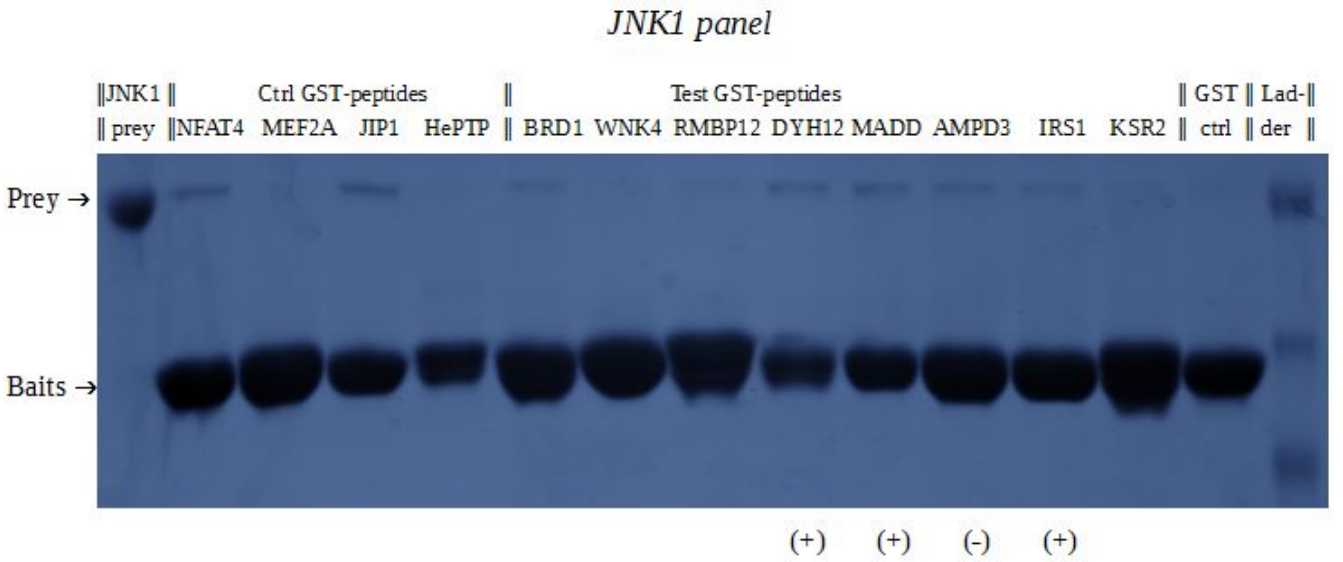
Figure 16: Workflow of the study, starting from the FoldX-based ranking.

Traditional binding-based assays are unsuitable to reliably identify novel D-motifs

After completion of initial lists, we chose a number of candidate proteins from each motif type to test. Testing proteins on a large scale immediately at full length, in a cellular environment was unfeasible. Due to the difficulty to reliably express full-length human proteins of large size (>1000 amino acids) in recombinant systems, we opted for a fragment-based approach. First, we attempted to set up assays based on binding alone. For the purpose of testing the method, I cloned a representative set of both known and potential D-motifs predicted with an earlier algorithm of ours (based on scoring matrices calculated from published examples alone, see Garai et al, 2012).[31] The short fragments corresponding to the D-motifs were inserted into the pETARA vector, endowing them with N-terminal GST and C-terminal His_6 tags. The complete set of 23 constructs, including a negative and several positive

controls, was expressed in *E. coli* cells and purified with Ni-NTA chromatography, then subjected to parallel **GST pull-downs**, with inactive (dephosphorylated) ERK2, JNK1 and p38α proteins. Unfortunately, it turned out that each MAPK behaves differently under the conditions of this assay. Moreover, although the assay itself was clearly reproducible, signal strength was often poor, especially with p38α (even after western blotting). But the greatest problem of this assay was the difficulty of threshold setting, with the "positives" and "negatives" often separated by marginal intensity differences only. The fact that it could not differentiate between binding through the D-site or spurious binding was another complicating factor (as later my colleague Ágnes Szonja Garai was able to confirm with fluorescence polarization based assays, some novel RevD-like motifs with "positive" signal did not interact with their predicted binding site at all). A sample GST pull-down experiment is shown on the figure below (as indicated, some motifs were later re-tested with other methods, giving a clearer picture of their true behaviour against the same MAPKs)





Figures (17): Coomassie-stained gels showing the outcome of an early pull-down experiment testing the MAPK-binding potency of various novel D-motifs. To illustrate the difficulty of threshold setting, the (+) and (-) signs indicate motifs that unambiguously tested negatively or positively in the dot-blot array experiments (see later). Unfortunately, western-blots resulted in similarly difficult-to-evaluate signal intensities (see supplemenary figures).

To remedy the situation, I also tried to apply a different strategy. In a previous article identifying several novel JNK-binding docking motifs, a solid-phase peptide array was used for screening. To test the applicability of the latter method, we ordered the synthesis of a D-motif test panel, from the same company and the same technology (cellulose-based hydrophilic support) as it was used by others for D-motif identification.[130] Since we knew that distance from the support might influence binding, the test set included the same motifs"shifted" N- or C-terminally (+ or -). We also tested the potential impact of peptide hydrophobicity on its exposure to the aqueous phase; and even included "synthesis controls", i.e. difficult-to synthesize peptides, due to their high proline content and direct Pro-Pro coupling. The test panels were used for binding assays, similar to a pull-down, then subjected to western blotting. Due to extremely low signal strengths (this was remedied by isotope labelling in previous articles), we had to modify the detection protocol, and use "tandem" western blotting (see materials and methods).

Unfortunately, the results obtained with the solid-phase peptide arrays were similarly disappointing as the earlier GST-pull downs. Although distance from the support did not appear to grossly influence binding, peptide hydrophobicity had a very pronounced effect, as did the difficulty of synthesis. The relatively hydrophobic motifs of Mcl1 and PDE4B - giving no signal here - were later shown to be

perfectly functional in dot-blot arrays (PDE4B has also been tested in FP titrations). The motifs of the closely related RHDF1 and RHDF2 proteins - with a very different signal here - gave similar signal and binding strength under both dot-blot arrays and FP titrations, pointing to low Pro-Pro coupling efficiency as a potential limiting factor of synthetic peptide arrays (generating the false image of "poor binding" with Pro-rich motifs). Also, the assay against p38 yielded higher binding on the non-cognate NFAT4 peptide than the cognate MEF2A, which would have been impossible, had the peptides been synthesized and exposed to solution to a similar degree.[31]

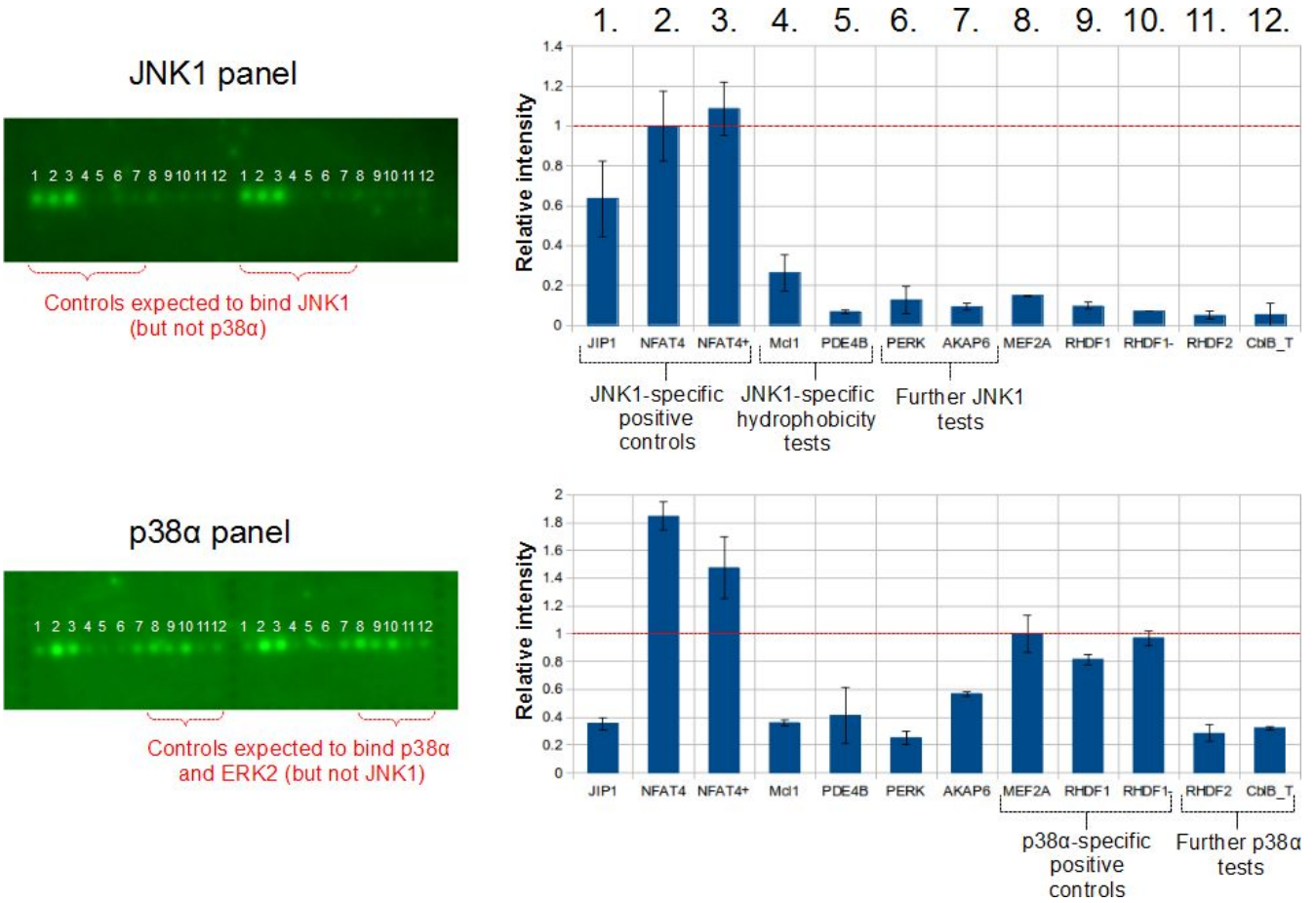


Figure 18: Testing of D-motifs on a solid-phase synthetic peptide array. JNK1 only gives signal with two well-known positive controls, but not with any other peptide. Peptides predicted to have poor water solubility show meager binding despite the hydrophilic support. Also, the profiles obtained for p38α contradict all previous results, and suggest inherently unequal epitope density on the glass slides.

Screening for docking motifs is possible with a novel solid-phase phosphorylation array

My former experiments showed that simple binding assays (such as pull-downs with recombinant D-motif containing proteins or immobilized solid phase peptide arrays) often lack the robustness to reliably detect low affinity (Kd ~ 1-10 micromolar) protein-peptide interactions. Therefore we decided to develop a different assay which was based on substrate phosphorylation enhancement on a solid-phase support. As the majority of known D-motif containing proteins are MAPK substrates, this is probably the nearest one can get to capture the original function of these motifs.

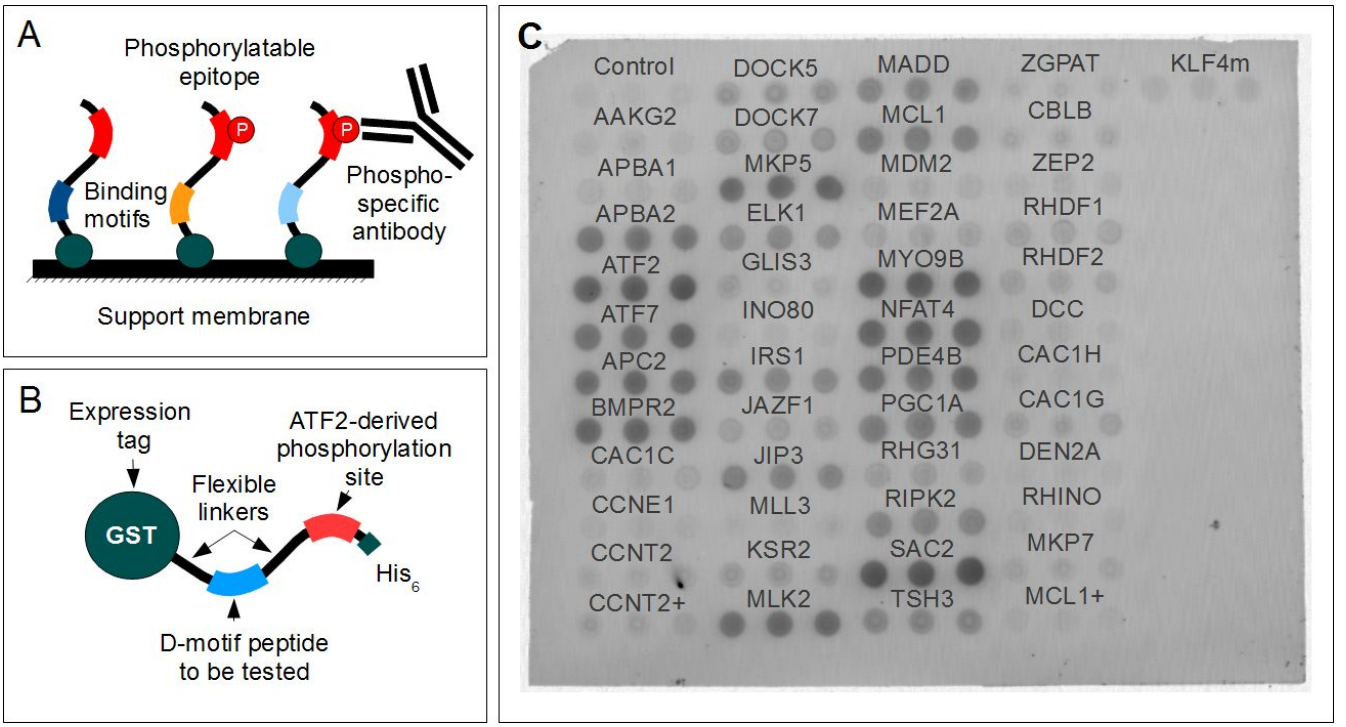


Figure 19: Principle of the dot-blot array (A), schematics of the sythetic substrate (B) and the ability of the method to capture MAPK substrate phosphorylation enhancement (C).

For this purpose, an artificial substrate was constructed containing the D-motifs as well as the Thr71 phosphorylation site from ATF2, which is a well-known MAPK target site. All D-motif encoding short segments were ligated into the same vector. As linkers and substrate sites in the recombinant proteins were identical, the "docking efficiency" of the given motifs could now be directly compared to each other. To make comparative experiments easier, I also developed a simple method to immobilize the

proteins on a solid support, and perform phosphorylation in a single reaction to make assay variability even lower. Testing of the assay indicated that phosphorylation of the reporter solely depended on the presence or absence of specific docking motifs and phosphorylation of the target site was always low without a functional D-motif (the target site alone did not determine specificity).

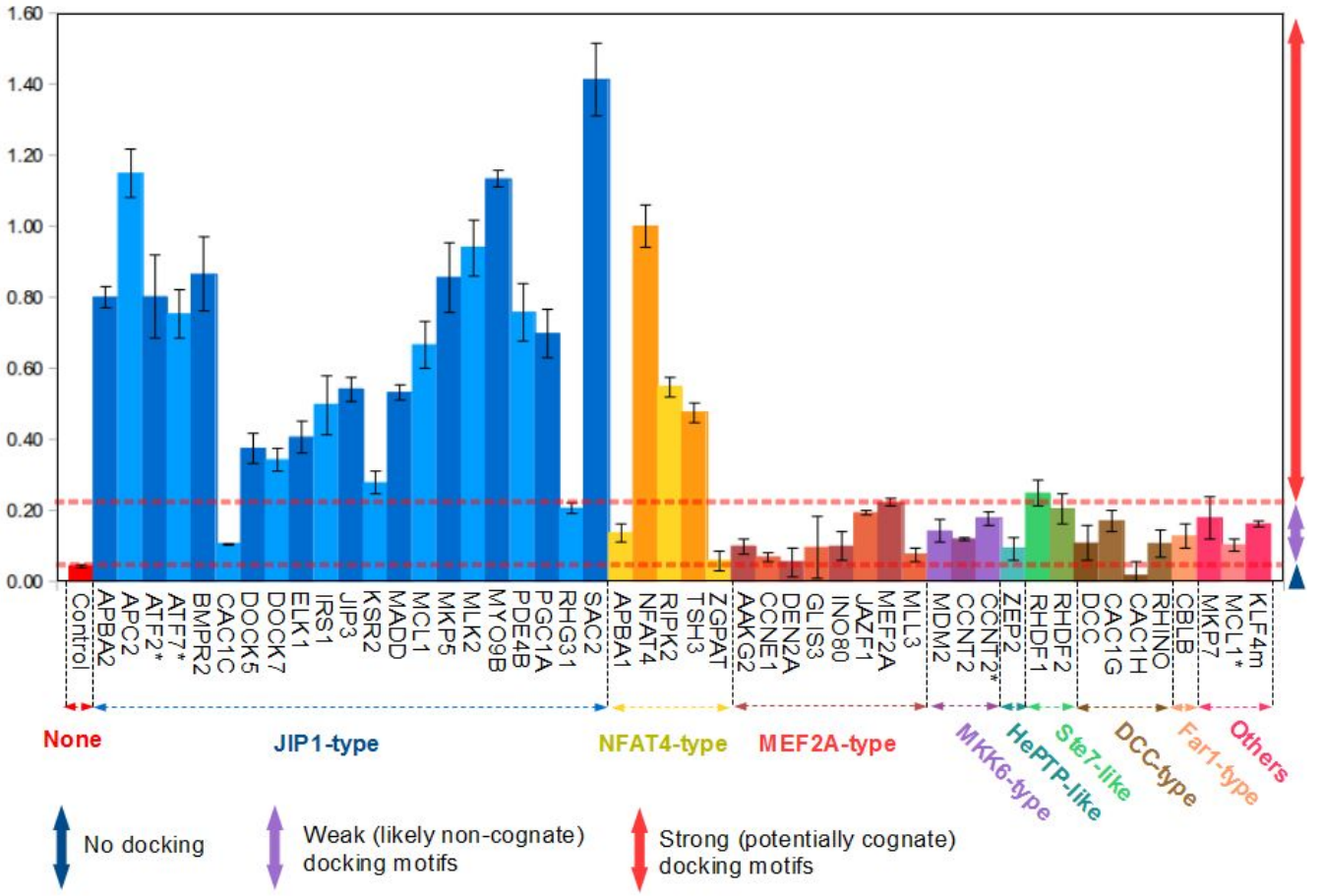


Figure 20: Evaluation of a dot-blot panel (with 48 constructs), showing phosphorylation enhancement values against JNK1. The non-cognate cutoff value was set to the one produced by MEF2A, known to be selective to ERK2 and p38α only in other settings as well as in vivo experiments.

In the final panels, we included 71 different constructs: 64 of these were directly selected from the lists, with a negative control added. We also included an additional 6 motifs based on sequence similarity to known motifs: This was done in order to test our algorithms; and to check if motif definitions were sufficiently inclusive or not. Out of 70, a total of 62 motifs were found to interact with at least one type of classical MAPK (ERK2, JNK1 or p38α). In particular, we were able to detect several novel interactors based on the JIP1, NFAT4, MEF2A, MKK6 and broader DCC models. As for our hypothetic

Ste7 model, we also found a novel hit: in the form of RHDF1 (and the closely related RHDF2) proteins. Such a high number of hits suggests that docking motifs are in fact quite widespread in the human proteome; With the right modelling and screening approaches, they could even be detected fairly reliably: Our screens included some additional, known positives as further controls (e.g. BMPR2, Elk1 and JIP3 for JIP1 types; JunD and MABP1 for NFAT4 types):[131–134] These additional controls were all shown to be functional, truely cognate motifs in the dot-blot screens.

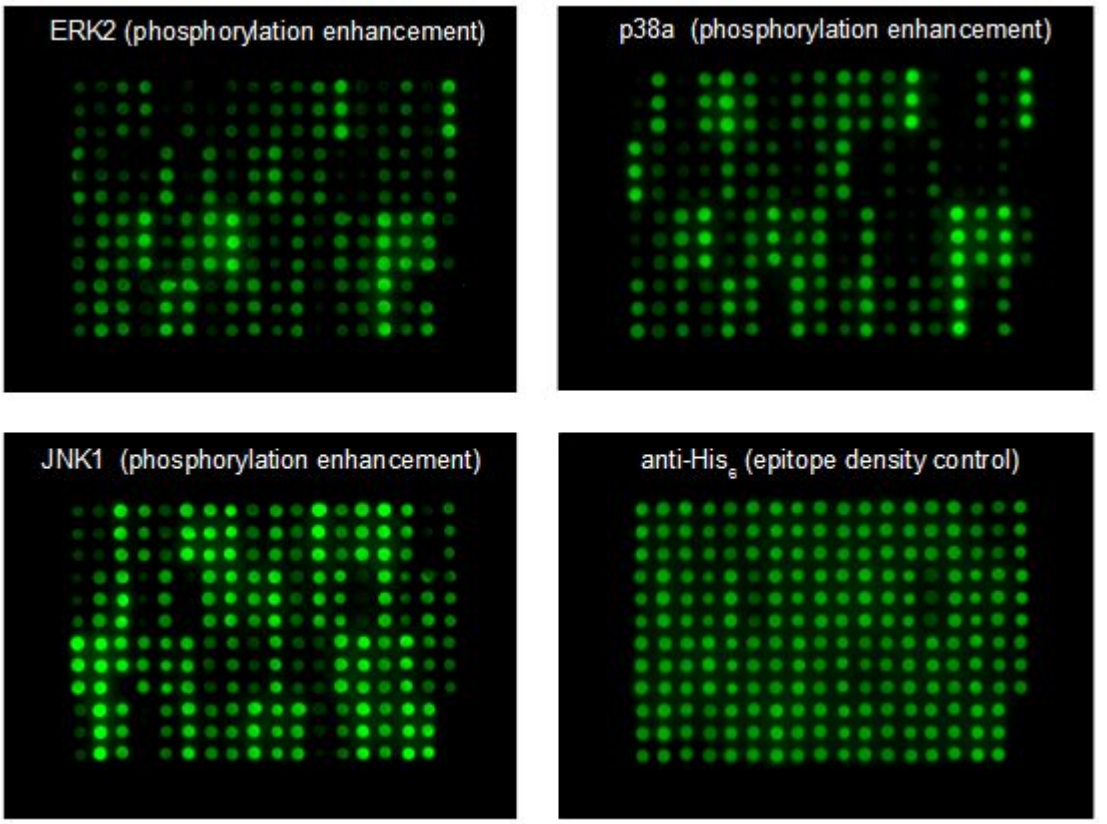


Figure 21: Dot-blot arrays in false color, with 71 different synthetic substrate constructs printed onto nitrocellulose membranes as triplicates, against three different, activated MAPKs as well as an epitope density control (using the C-terminal His_6-tag to assess protein integrity). [see supplementary files for detailed evaluation]

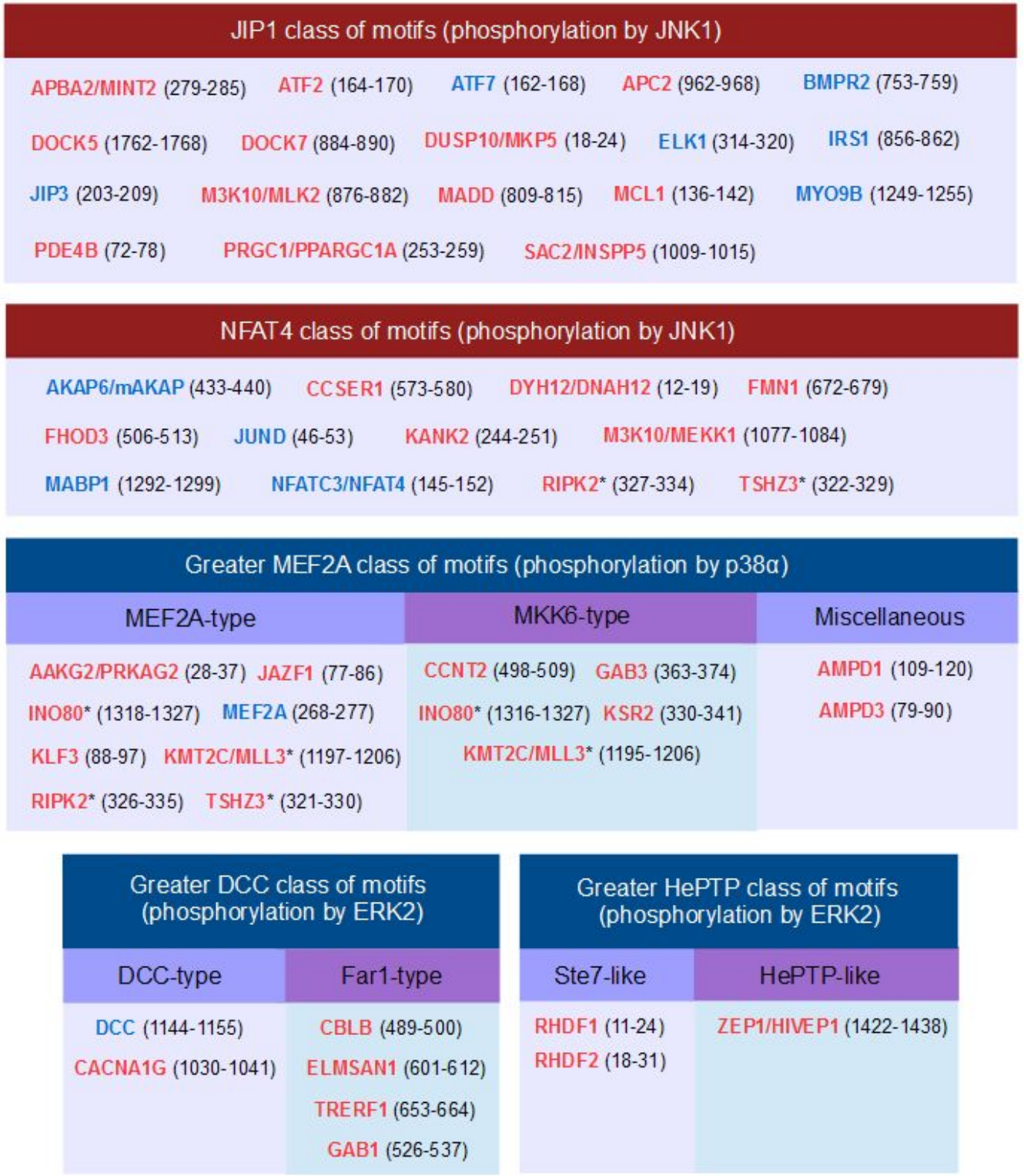


Figure 22 List of already known (blue) and novel (red) D-motifs that were identified as functional, based on the dot-blot array experiments. The numbering corresponds to the position of the motif in the "base" isoform indicated by UniProt. In the case of proteins, more than one synonymous names are given. Asterisks are appended to the names of motifs belonging to multiple classes simultaneously.

Fluorescence polarization assays and MAPK profiling

To show that the phosphorylation enhancements were indeed due to the presence of canonical MAPK docking motifs, we tested 14 of our novel motifs in fluorescent polarization (FP) titrations against well-known D-motifs. Here, short peptides representing a novel docking motif were titrated in vitro against fluorescently labelled controls bound to a MAPK. Anisotropy recordings were used to draw competition curves and calculate their exact dissociation constant (Kd). These measurements were principally carried out by Ágnes Szonja Garai and Klára Kirsch (see supplementary for raw data).

Protein name	Peptide sequence	Kd values (μM)			
		ERK2	JNK1	p38α	
APBA2	RHEARPKSLNLL	-	10.8	-	JIP1 class
ATF7~	EEPTIVRPGSLPLHLGYE	-	21.1	-	
DOCK5	KAQRPKSLQLM	31.8	5.1	-	
DOCK7	RSAVRPASLNLNRSR	51.2	7.3	-	
IRS1	RLARPTRLSL	-	3.5	74.6	
MYO9B	LERPTSLALD	-	15.9	-	
MKP5	EESRPVRPQDLNLSLDSE	84.7	19.3	-	
PDE4B	EGDGISRPTTLPLTTLPS	-	16.4	-	
DCX	SLRKHKVDLYLPLSL	59.0	16.7	29.5	NFAT4 class
AAKG2	SQKRRSLRVHIP	17.5	-	8.2	greater MEF2A class
CCNT2	KKEKSGSLKLRIPIP	14.4	-	24.5	
GAB3	SLRHRDKRLSLNLPSR	22.6	14.5	5.2	
JAZF1	SLKKKIQPKLSLTLS	15.0	-	8.9	
KSR2	KKKSKPLNLKI	13.5	24.7	4.9	
GAB3+	KPQRKSRPPPLDLRNLS	0.7	13.2	2.0	gtr. DCC class
RHDF1	SLQRKKPPWLKLDIPS	0.9	-	1.2	gtr. HePTP class

Figure 23: Results of competitive fluorescence polarization (FP) titrations with peptides representing several novel D-motifs. Dissociation constants confirm grouping of motifs into two major clusters (primarily JNK1 binders vs primarily ERK2/p38α binders), suggested by their structural models (on the right). (Keys: ~: sequence with non-native flanking amino acids. -: Kd is above limit of quantitation (>100μm) +: motif based on homology to Gab1, not directly tested in the dot-blot screens)

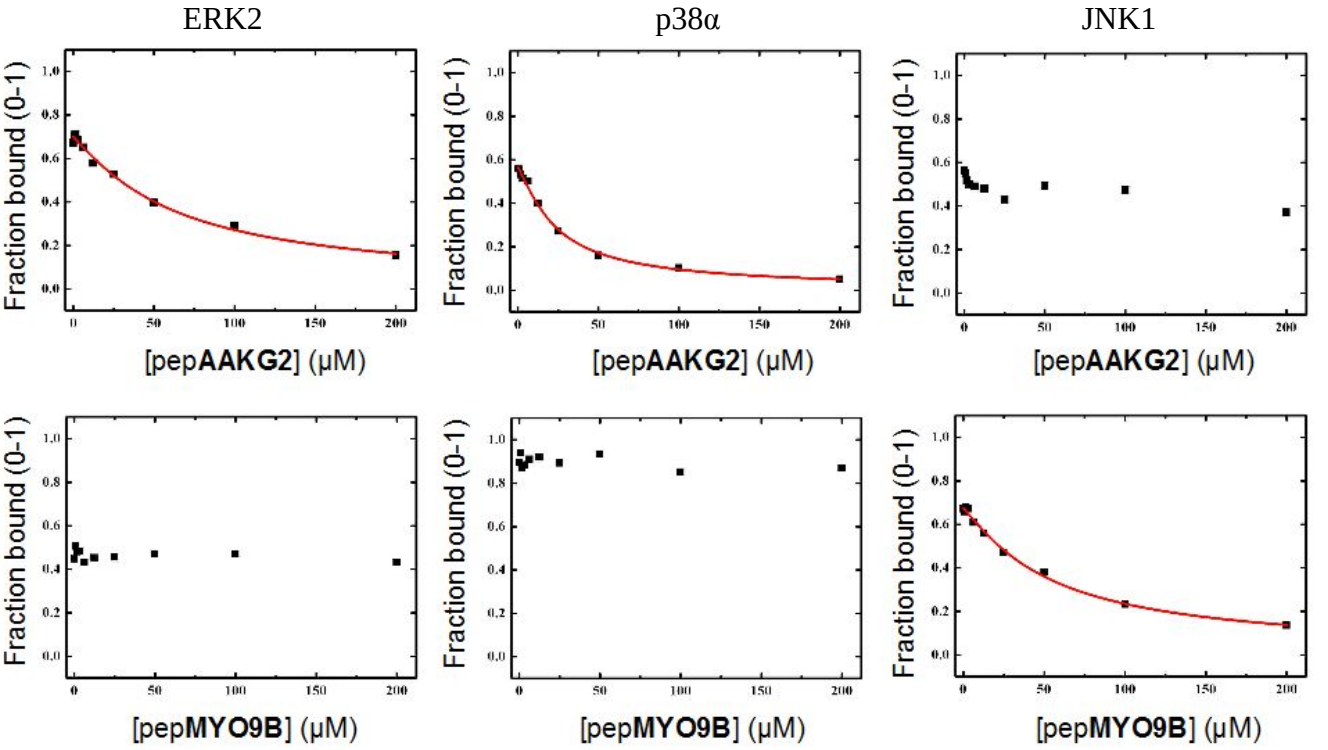


Figure 24: examples of competitive FP titration curves, obtained with a peptide selective for ERK2/p38α (AAKG2), or with a JNK1-specific one (MYO9B), against fluorescently-labelled competitors (see materials and methods for details).

Binding affinities obtained by FP titrations also allowed us to examine the specificity profiles of D-motifs. The tested peptides could be clustered into two groups, based on their sequences and affinities. Similarly to earlier results, these experiments confirmed the strong correlation between the ERK2 and p38α binding ability of a given motif and binding results also reflected the fundamental lack of correlation between ERK2/p38α and JNK1 association. These observations did agree well with phosphorylation enhancement results from dot-blots. There was no positive correlation between the profiles of the JNK1 vs. p38α or the JNK1 vs. ERK2 pairs (Pearson's r=0.003 and r=-0.280, respectively). At the same time, a modest correlation was observed between ERK2 and p38α (r=0.680). This MAPK profiling confirmed our structural models. Practically no strong JNK1-binding motif was found from other than the JIP1- or NFAT4-type classes. Most novel p38α interactors, on the other hand, belonged to the MEF2A-, MKK6- or DCC-types as expected. The profiles recorded for ERK2 were somewhat eccentric: Apart from sporadic interactions with JIP1-type motifs (including Elk1, that is well-known but not predicted by any structural model), ERK2 interacted with most short, MEF2A and MKK6 type motifs weaker than p38α. The truely strong ERK2 interactors fell into the broader DCC class or were long, helical motifs (RHDF1and RHDF2 that are Ste7-type and ZEP1 that is presumably HePTP-like, but with a longer linker).

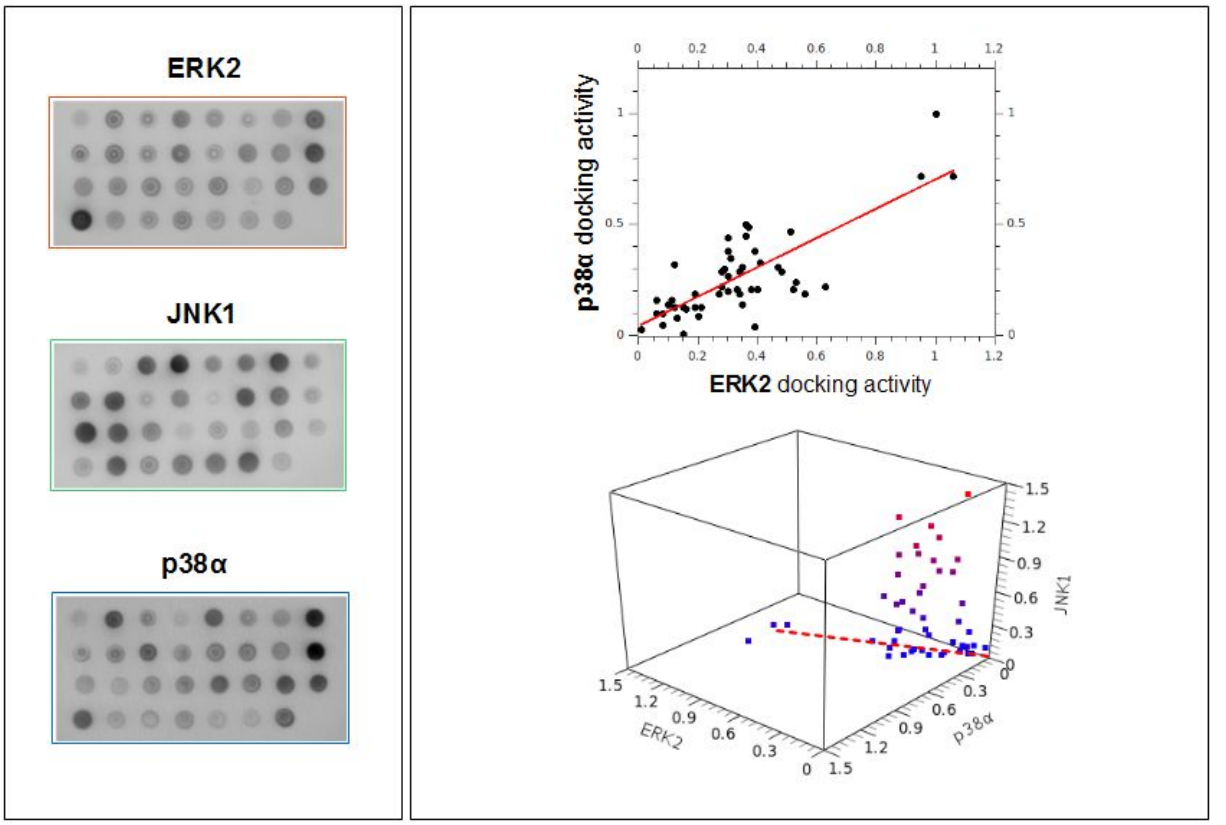


Figure 25: The D-motif preference patterns of ERK2 and p38α are similar in dot-blot arrays but very different from JNK1 (left). By the same virtue, the correlation of ERK2 and p38α docking is more-or-less obvious in pairwise comparisons of docking enhancements. The results of a 48-membered partial panel are shown on the right in a 3D plot; correlation is indicated by a regression line between enhancements of ERK2 or p38α catalytic activity (on a relative scale set to MEF2A and NFAT4).

Validation of interactions in living cells

To test whether the docking motifs were also functional in their native context, we set up a fluorescent protein fragment complementation (BiFC) assay. This cell-level testing were done by Anita Alexa. In this series of experiments, one fragment of YFP was fused to either ERK2, JNK1 or p38α. The other fragment was joined to our protein of interest; and we measured the intensity of fluorescence after co-expressing the two fragments in HEK 239 cells. As it is well known, fluorescence intensity not only depends on complementation efficiency, but also on protein concentrations. To enforce equal concentrations, we always monitored the level of protein expression with western blotting. However, the degree of complementation between a given MAPK and different partner proteins (even with

similarly sized or arranged constructs) could still vary. Thus all BiFC signals were compared to the results obtained with the same construct but lacking the docking motif.

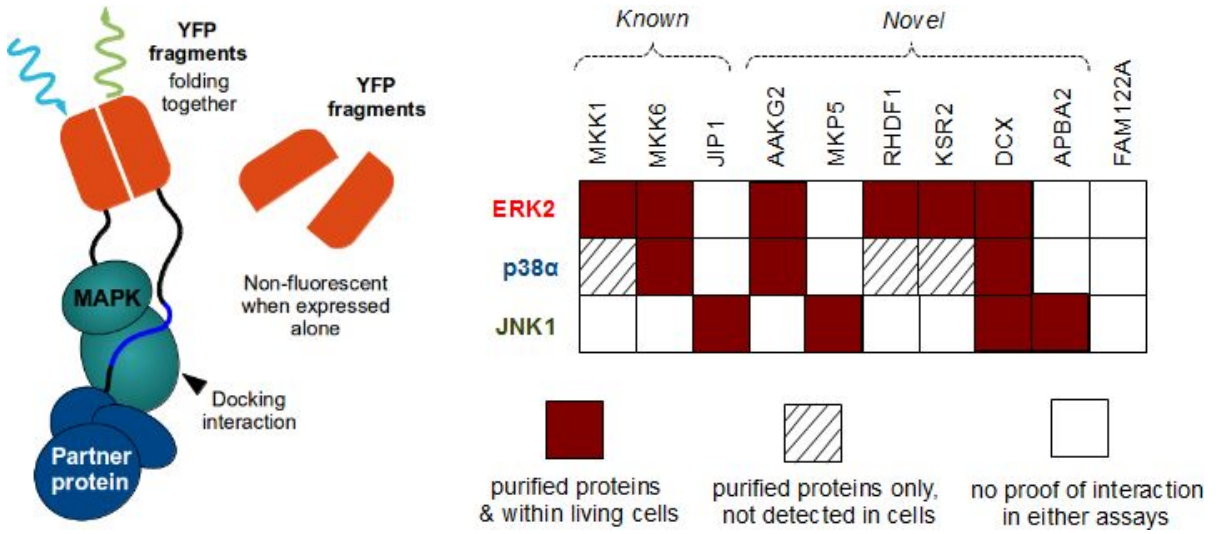


Figure 26: The schematics of a BiFC experiment (left) and the in vitro found interactions detected in our cell-based assay (right).

Well-known MAPK partners, such as MKK1, JIP1 or MKK6 display a pattern consistent with the specificities of their D-motifs (see above). Such interactions are also greatly diminished or abrogated after the loss of the docking motif. The same is true for several of our novel MAPK partners. For example, the **AMP-activated protein kinase subunit γ2** (AAKG2) displayed clear signs of interaction with p38α (and ERK2), but not with JNK1. However, AAKG2 is known to admit multiple, shorter isoforms through use of alternative initiation codons. One such variant (isoform C) is only shorter by 44 amino acids. This natural deletion mutant lacking the N-terminal (MEF2A-type) docking motif showed a greatly reduced level of fluorescence for both partners. The differences in fluorescence were readily visible on cells under a fluorescent microscope. All intensities - as well as their reduction in the mutants were also comparable with the ones observed in control experiments. These results are well in line with our in silico predictions and in vitro fragment-based experiments. To show that the case of AAKG2 is not unique, we also validated a number of other interactions in cells, using full-length proteins whenever possible.

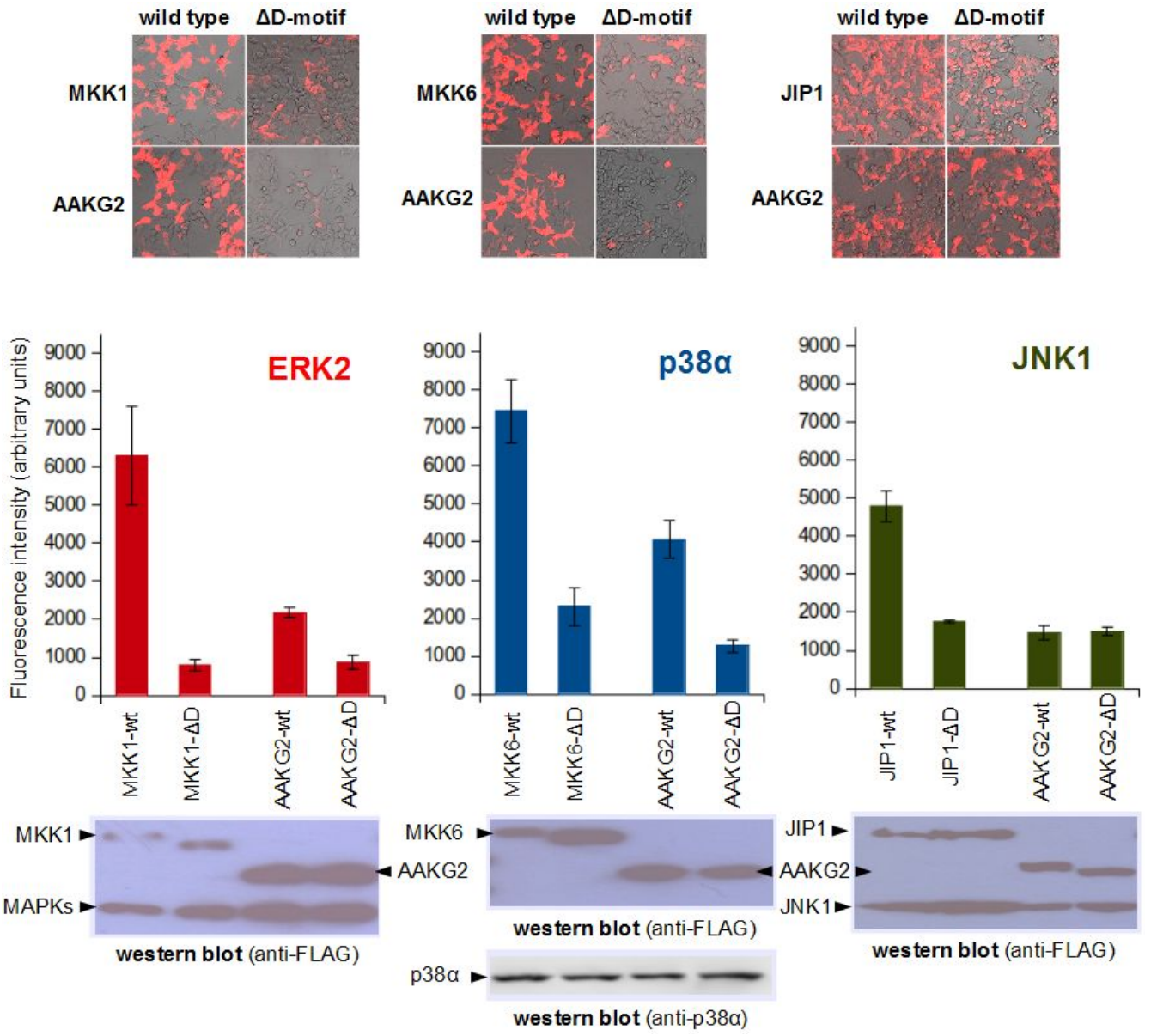


Figure 27: Experiments showing the D-motif dependent in vivo interaction between the AAKG2-ERK2 and the AAKG2-p38α pair, and the lack of interaction with JNK1 through the same motif. Moreover, the strength of interaction appears to be comparable to well-known partners. The raw image of cells under a fluorescence microscope is shown on the upper panels, while the lower panels show recordings on a plate reader (typically N=6) as well as a western blot from pooled samples. All visible differences were significant (ERK2/MKK1: $p=0.001$, p38α/MKK6: $p=9.6*10^{-6}$, JNK1/JIP1: $p=3.3*10^{-3}$, AAKG2/ERK2: $p=2.1*10^{-5}$, AAKG2/p38α: $p=2.8*10^{-5}$, but AAKG2/JNK1: N.S. on a two-tailed Welch's T-test).

The **Inactive rhomboid protease 1** (iRhom1 or RHDF1), which is known to regulate the secretion of various paracrine factors (TNF-α, TGF-α, amphiregulin) through ADAM17/TACE, was found to

harbor a Ste7-like docking motif at its N-terminus.[135,136] Such long motifs can provide specific and high affinity interactions. Despite its low expression, a considerable fluorescence signal was seen when full-length RHDF1 was co-transfected with ERK2 and this was greatly reduced when mutant RHDF1 was used that lacked its D-motif. A further protein harboring ERK2 and/or p38α interacting motifs, **Kinase suppressor of Ras 2** (KSR2) could not be expressed in full length. However, we succeeded in expressing a shorter fragment of KSR2, that encompasses the entire region unique to KSR2. (This segment is a product of a single exon, not found in KSR1, and was predicted to be fully disordered). This KSR2 fragment readily interacted with ERK2, judged by the reduction of signal intensity when a mutant protein was used (see the figure on the previous page).

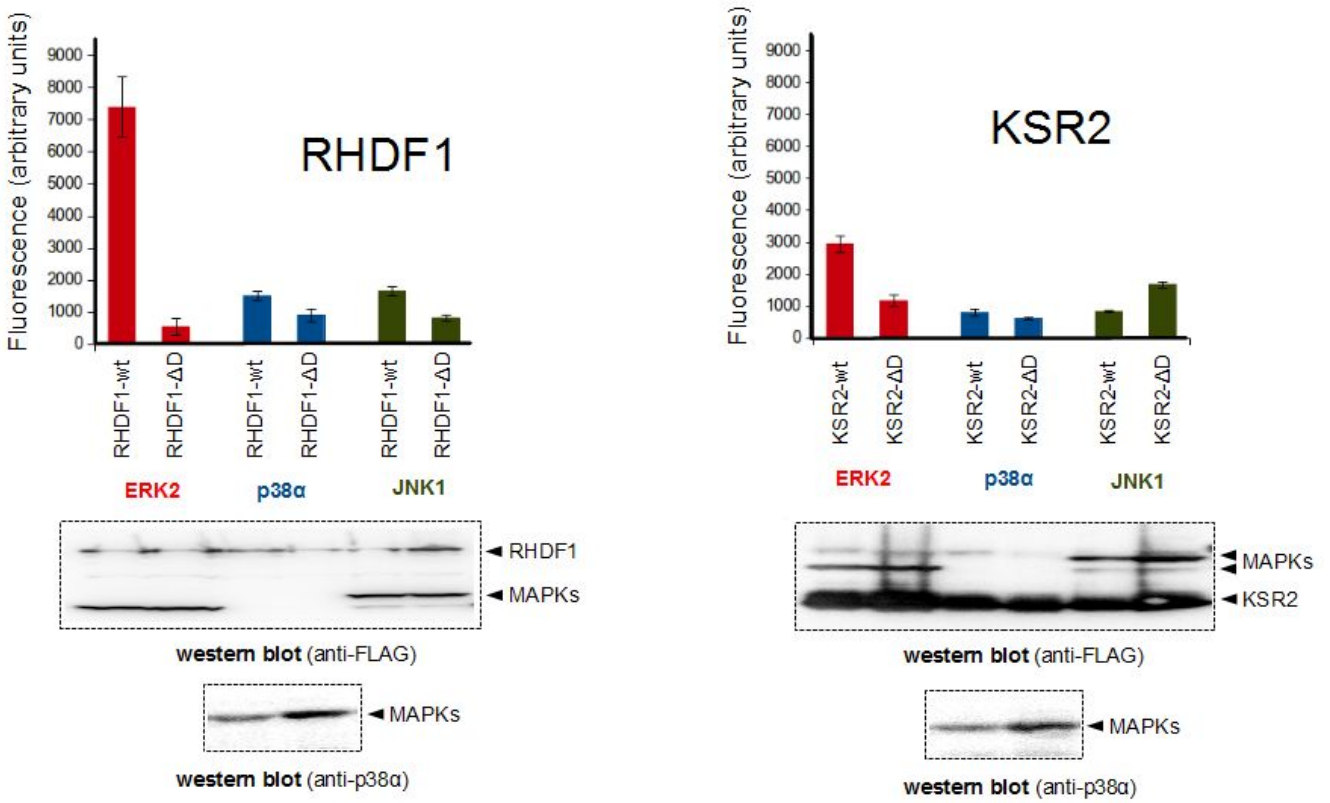


Figure 28: BiFC assays showing D-motif dependent interactions of ERK2 towards the transmembrane growth factor shedding regulator RHDF1 as well as the ERK1/2-pathway auxiliary factor KSR2. The wild-type versus delta-D-motif mutant differences were all strongly significant (RHDF1/ERK2: $p=3.0*10^{-4}$, KSR2/ERK2: $p=9.6*10^{-6}$ on a two-tailed Welch's T-test assuming unequal variances)

A number of JNK interactors were also tested.. **MAP kinase phosphatase 5** (MKP5, also known as DUSP10) is known to dephosphorylate both p38α and JNK1.[137] The structural basis of its specific interaction towards p38α has already been revealed.[138] In contrast, the domain(s) or motif(s) that allow

MKP5 to recognize JNK1 were never identified. Our proteome wide search for JNK binding D-motifs suggested that MKP5 contains a functional JIP1-type motif: This could selectively mediate its binding to JNK. In accordance with that model, MKP5 showed a high degree of fluorescence when co-expressed with JNK1. The same signal was greatly reduced when the N-terminal JIP1-type docking motif was removed. The X-chromosome linked lissencephalia protein **Doublecortin-X** (DCX) has also proven to be an important example of D-motif-dependent interactions (in contrast to earlier publications).[139] We identified an alternatively spliced isoform of DCX that harbors a classical NFAT4-type docking motif. (The previously described isoform also contains an atypical, loosely NFAT4-like motif at this place.) DCX interacted with JNK1, but BiFC signal was greatly reduced when the C-terminal DCX docking motif was absent. Interestingly, strong reductions were seen with ERK2 and p38α as well. These findings can be explained by the capacity of the DCX C-terminal motif to interact with all three MAPKs to a limited extent, under FP titrations as well as dot-blot arrays.

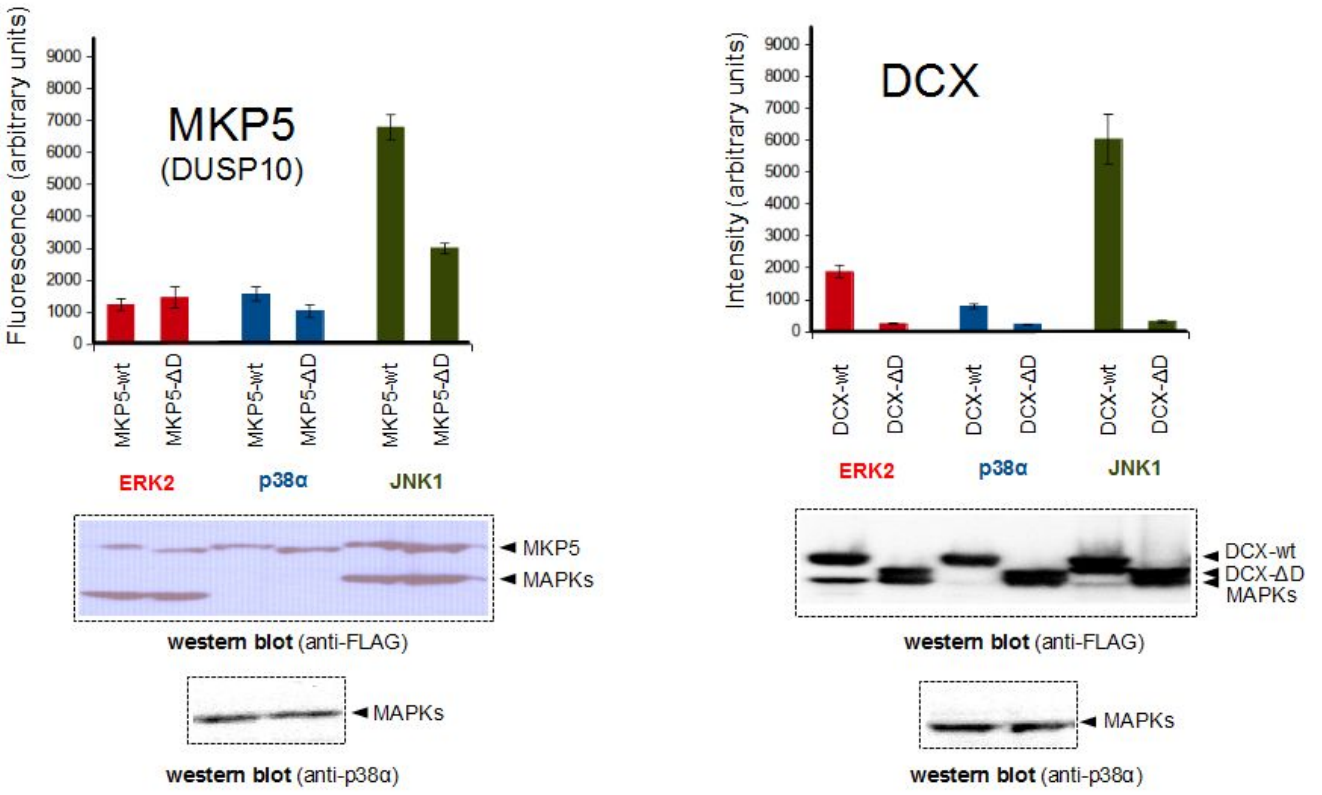


Figure 29: BiFC experiments demonstrating D-motif dependent interactions of JNK1 with the MAPK phosphatase MKP5 and the microtubule-binding protein DCX (note that the latter also seem to interact with other MAPKs, but to a much lower extent). All interactions mentioned here were significant. (MKP5/JNK1: $p=1.7*10^{-7,}$ DCX/JNK: p=0.0006, DCX/ERK2: p=0.0006, DCX/p38α: p=0.0007 with a two-tailed Welch's T-test)

Another JNK interacting protein gave low, but still detectable signals. The **synaptic adaptor protein MINT2** (or APBA2) was found to harbor a JIP1-type motif. APBA2 is an exclusively neuronal protein with a domain architecture remotely similar to JIP1 itself.[140] The fluorescent signal was the highest when this protein was co-expressed with JNK1 and this was detectably reduced with a mutant which lacked the D-motif. Despite small effect size, this change was still much larger than the relatively meager changes observed for p38 or ERK2. Unfortunately, fragment complementation efficiency also depends on the relative orientation of fragments. Thus BiFC can not be used for quantitative analysis as a low fluorescence intensities often do not mean weak interaction. The fluorescence signal would only be proportional to binding affinity if all constructs were completely identical in length as well as in structure. To prove that results of full-protein based BiFC assays still correlate with the short motif based methods to a certain extent, we also introduced a true **negative control**: the NFAT4-type D-motif from the functionally unknown FAM122A protein gave close to zero signal in dot-blot experiments. In accordance with this, the full-length FAM122A gave very low signal in fragment complementation assays, and importantly, removal of this N-terminal pseudo-motif did not change complementation efficiency, as expected.

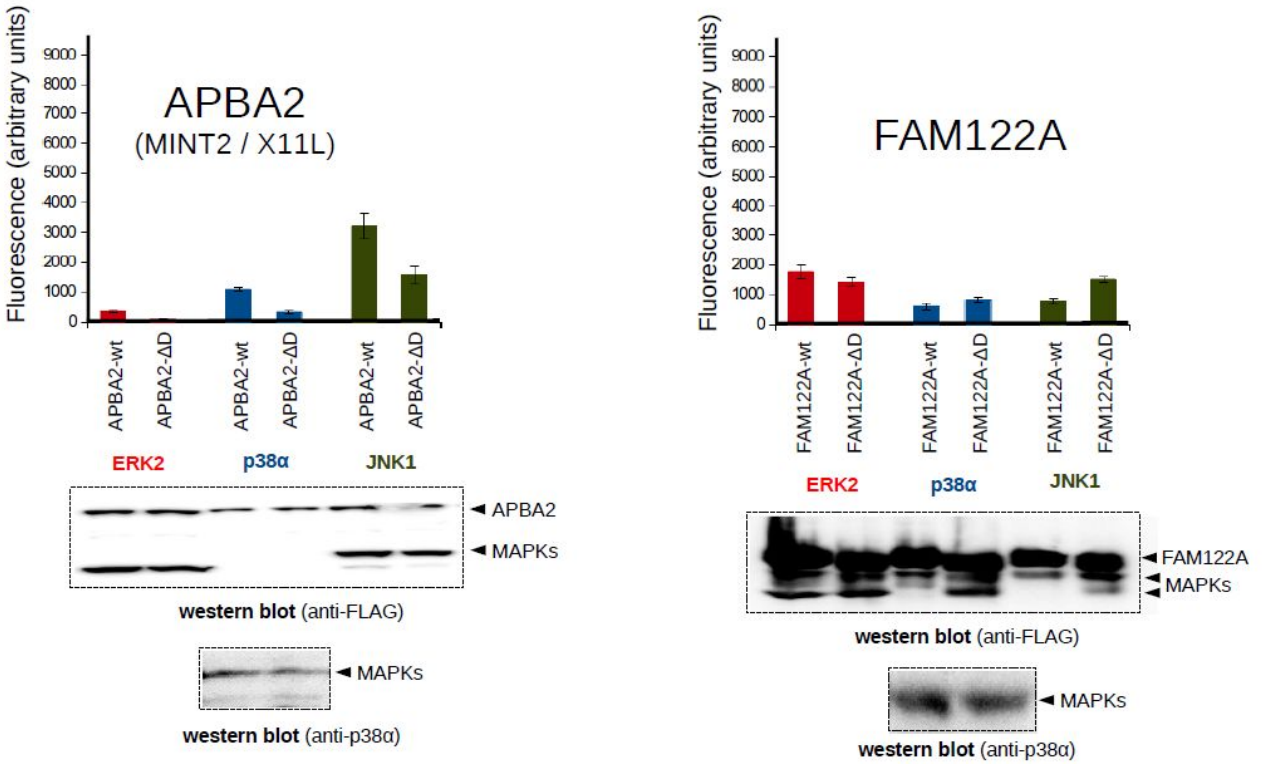


Figure 30: BifC experiment with APBA2 and FAM122A proteins as well as their appropriate mutants, demonstrating a D-motif dependent interaction (with possibly multiple MAPKs, not just JNK1, with all differences significant) in the case of APBA2, and the lack thereof in the case of FAM122A

Refinement of structural models and PSSM building

As a conclusion to the experiments, we could use the results to improve earlier structural models. Evolutionary conservation of certain motifs was also a great tool to examine sequence conservation or diversity per each position. We concluded that some of our initial motif definitions were too loose, especially the JNK-binding motifs. In the end, the **JIP1-type motifs** could be re-defined as [RK]P[^P][^P]Lx[LIVMF]. The removal of prolines from several positions as allowed amino acids was also supported by previous studies. The leucine, on the other hand, was typically extremely conserved at the middle pocket (similar results were obtained for the equivalent position of the NFAT-type motifs). The strict **NFAT4-type consensus** could be written as [RK][^P][^P][LIM]xLx[LIVMF]. Here, the exclusion of prolines is based on their structural incompatibility with the helical segment and also supported by earlier experiments. Notably, some rare JNK interactors do defy this rule (most obviously, the D2 motif of MKK7), thus other, neither-NFAT-nor-JIP binding modes must still exist. Nevertheless, it was recently shown that the D2 motif of MKK7, too, mimics the arrangement of either JIP1 or NFAT4-type motifs, although somewhat imperfectly.[141]

On the other hand, several of our definitions on p38α and ERK2 interactor motifs turned out to be too restrictive. We could easily find examples (such as the motifs of AMPD1 or AMPD3) that did not match either the **MEF2A- or MKK6-type arrangements** on their N-termini, but were nevertheless valid interactors. For the sake of completeness, we had to abandon the tight sub-classification, and use a global, loose motif definition for the greater **MEF2A class:** [RK]x{2,3,4}[LIVMP]-[LIV]-[LIVMF]. Here the only improvement involved the rightmost (B) pocket - evolutionary homologs of our hits show that proline is normally not found here, only in truely exceptional cases. Though the **Ste7-type** consensus was also proven as a valid, existing class by a novel example: RHDF1, comparison of evolutionarily related proteins (RHDF1 and RHDF2) suggested that the first charged **θ** position is merely optional here (this amino acid likely also faces the MAPK surface at an unfavourable angle). We could find an example of further motifs (from ZEP1) that would presumably fit a broader **HePTP-like** model (with helical N-terminus). Yet it could not be predicted by systematic searches due to its divergent linker length (from **θ** to $\boldsymbol{\varphi}_L$), 6 amino acids in the case of ZEP1 in contrast to 5 observed in all HePTP-related motifs. The greater **DCC class** - on the other hand - was represented by multiple new hits. Here our sequence-based predictions were considerably improved upon introduction of an additional constraint: having at least one proline amino acid in the $\boldsymbol{\varphi}_L$- $\boldsymbol{\varphi}_A$ linker. Several novel

members showed a conserved proline immediately preceding the φ_A position, resembling the motif from the yeast Far1 protein. This is in contrast to the other option seen in DCC and MEK2, where a proline is inserted after the φ_L position instead. But these are by no means mutually exclusive solutions: Even the evolutionarily related examples (comparison of the corresponding docking motifs of human Gab1 and Gab3 or the yeast Dig1 to Dig2) suggest frequent interconversion.

Once the consensus sequences were improved, we set out to build a **sequence-based method** to enable direct search for MAPK-interacting proteins from the human proteome. This was performed mainly with the help of Tomas Bastys (MPI, Saarbrücken). To increase the sequence space coverage, we included more than just the (known or novel) human MAPK docking motifs. A method was devised to use evolutionarily weighted sequences for each independently evolved (or sufficiently unique) motif. For this purpose, alignments were built from vertebrate sequences obtained by BLAST searches. Based on the refined consensus, motifs were classified as either potentially functional or non-functional. The motifs deemed "potentially functional" were re-aligned (with no gaps allowed) to the original sequence. In the end, the sequences were weighted by their distance from each other (alongside a distance tree) and the final frequencies were obtained by summing up all independent groups with equal weights. The resulting position specific scoring matrices (PSSMs) were constructed from full sets of evolutionarily related docking motifs (vetted based on the strict consensus). Such matrices have proven to be an efficient tool to identify MAPK-interacting proteins. High area under curve (AUC) values of the receiver operation characteristic (ROC) curve from a stringent five-fold cross validation imply an adequate coverage of motifs in the JIP1, NFAT4 and greater MEF2A classes: 0.98, 0.94, and 0.97, respectively. The correlation with the original, FoldX-based rankings was modest, but clearly present in the case of JIP1-type ($r = -0.62$) and NFAT4-type ($r = -0.59$) motifs. It was lower for the DCC ($r = -0.40$), MEF2A ($r = -0.30$) and MKK6 ($r = -0.26$) models, as somewhat expected, since the structural templates of these were incomplete (as the structures of the charged N-termini of some D-motifs are not known). Unfortunately, the lack of sufficiently diverse hits among DCC and HePTP-type motifs made PSSM construction impractical. A PSSM was still built for the greater DCC class, but only to compare it to the other three (see figure).

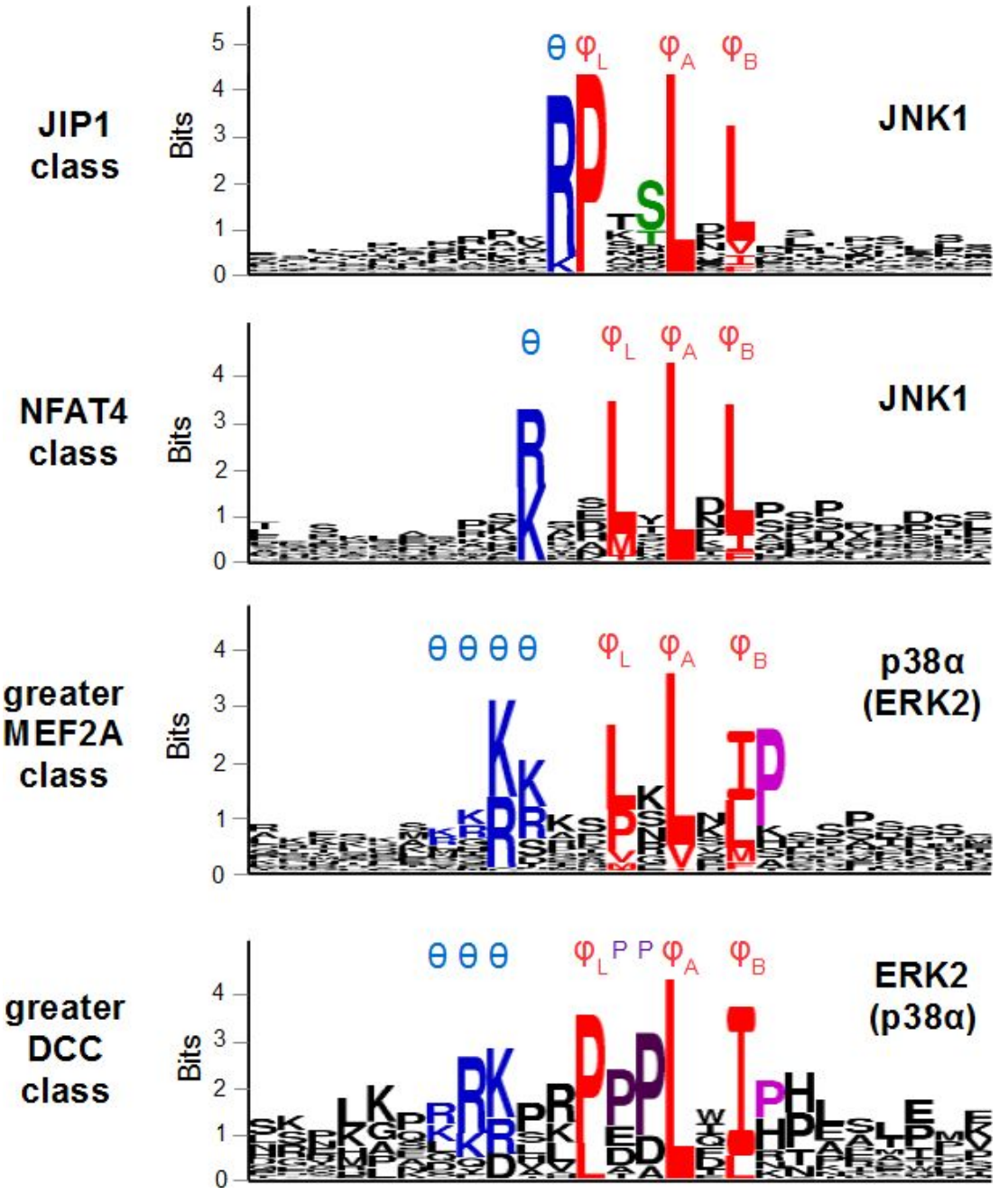


Figure 31: Graphic representation of the PSSM matrices for four major D-motif classes.Residues corresponding to the theta position are shown in blue, while the phi positions are red.

The resulting PSSMs were also checked from another perspective. Several examples illustrated that positional preferences could all be explainable on a **structural basis**. For example, the JIP1-type motifs showed a preference for serine or threonine in an X position preceding φ_L. From the JIP1-JNK1 structure, it is clear that this residue can form a hydrogen bond with an arginine on the surface of JNK1.[114] This observation explains the selection for amino acids capable to form hydrogen bonds. On the other hand, the p38α-binding motifs, especially those of the greater MEF2A class, displayed a clear

preference for proline in the position following the last hydrophobic contact point φ_B. The structure of MEF2A-p38α complex indicates that this proline forms an auxiliary hydrophobic interaction with the surface of the MAPK.[116] A slight preference for proline after φ_B was also seen in the greater DCC class. However, DCC-type peptides have a much more direct requirement for prolines in the segment between φ_L and φ_A as already noted before. The latter preference is likely tied to the requirement for prolines in order to maintain a type II polyproline helix.

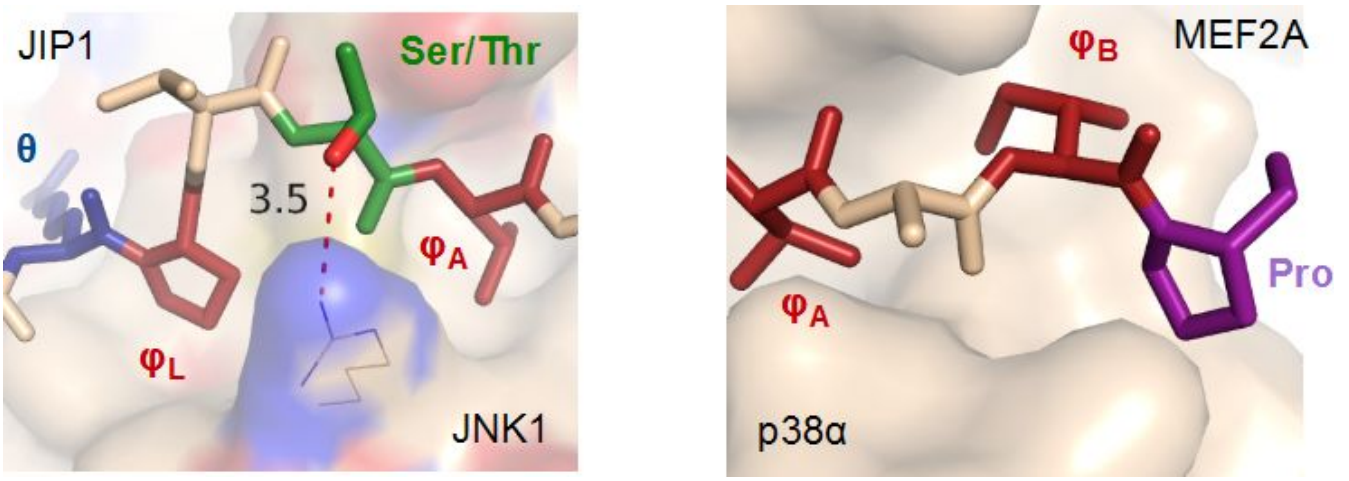


Figure 32: explanation for the non-core postion preferences in the case of JIP1 class (left) showing selection for residues capable of H-bonding and for the greater MEF2A class (right), frequently displaying Pro after the φ_B position, that can contact an additional hydrophobic surface.

A machine learning algorithm called **D-Finder** had been used to screen for JNK-associating proteins earlier.[130] This method applied a window based scanning procedure, selected peptides with a loose motif consensus (**φ-x-φ)**, and scored them using a trained hidden Markov model (HMM). We tested D-finder and compared its performance to that of the method described here. The resulting AUCs for JIP1-type, NFAT4-type and greater MEF2A-type motifs were 0.87, 0.75 and 0.60, respectively. Lower overall ROC performance in two-fold cross validation tests indicates that our method supersedes D-finder, with the latter performing not far above a random classifier in the greater MEF2A class. These results may reflect the fact that D-Finder was developed to identify JNK-binding motifs (JIP1 and NFAT4-types) alone. Within the greater MEF2A class, there is a greater variability in the number of positively charged residues and in linker lengths, and this structural diversity is better handled in our new method.

Predicted MAPK interactomes & novel pathways

Construction of robust and structurally meaningful PSSM matrices enabled us to search for MAPK docking motifs directly from the human proteome. Based on the distribution of our experimentally-validated examples on the lists, the best 100 hits were judged as the high-confidence set of predicted interactors. As it includes a rather large number of proteins that have little or no formal Gene Ontology (GO) annotation, we annotated all hits manually. For each protein, two terms were assigned, one "protein level" (low level) and one "systems level" (high level). Then all proteins were clustered to yield meaningful categories. Out of the three classes examined, the **JIP1**-types had the highest number of validated hits by far. Thus its predictions were deemed the most reliable, shedding some light on the interactome of JNK1. Interestingly, the majority of JNK-associating proteins (both experimentally validated and predicted ones) seem to be involved in cytoskeletal regulation. We encountered numerous actin-binding or microtubule-binding proteins, molecular motors as well as small G-protein partners. Docking motifs were even found on proteins localized to centrosomes, basal bodies, or those involved in the formation of primary cilia. Several other high-scoring hits suggest that JNK is intimately involved in the regulation of endo- and exocytosis. Among the less surprising categories discovered were the MAPK pathway components themselves (especially at the MAPK kinase kinase [MAP3K] level, as potential feedback elements), several transcription factors and other gene expression regulatory systems, or various ubiquitin ligases. In addition to adaptation to stress or apoptosis, JNK also appears to play an important role in embryonic development: JIP1-type docking motifs were also detected among components of the canonical Wnt pathway. (See the figure on the next page, showing all non-orphan categories that resulted from the clustering of hits.)

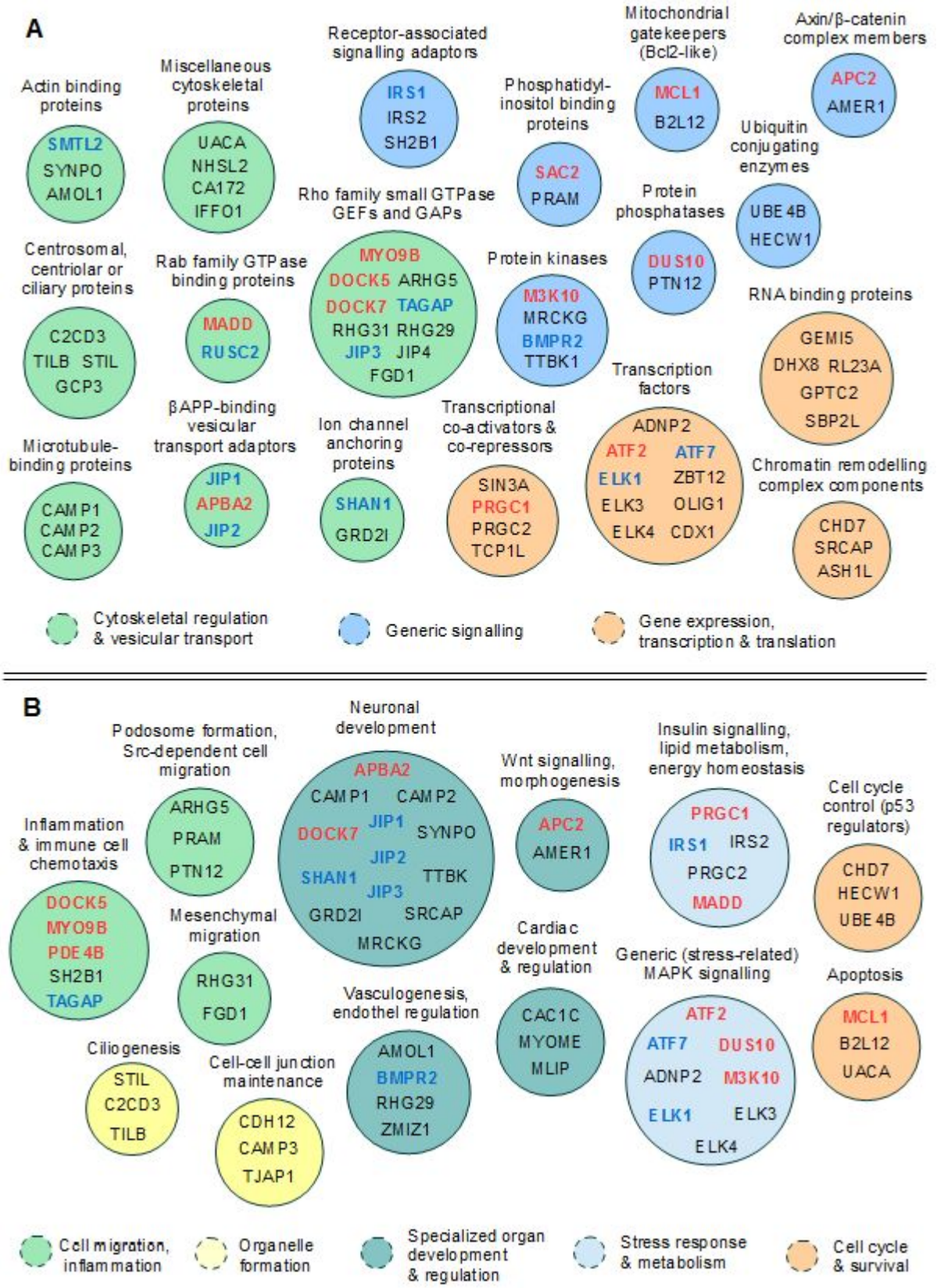


Figure 33: Low level (A) and high level (B) classification of the proteins with the best 100 JIP1-type motifs. Already-known partners are written in blue, while newfound interactors are red.

A considerable number of experimentally tested or predicted JNK-interacting proteins have preferentially or exclusively **neuronal functions**. We predict that the axons, nerve terminals and dendrites contain a high number of specialized JNK-interacting proteins; as do developing neuroblasts and their axonal growth cones. The inner cytoskeleton (microtubules and actin filaments) of neuronal cells appears to be the main target of JNK binders. Apart from JIP1, WDR62 and DCX (that are known to be involved in vesicular transport, microtubule organization and axonal growth, respectively), most of the interactions are fairly novel.[139,142,143] These newfound partners play a critical role in neuroblast fate commitment (Myt1L, SMARCD3), glial differentiation (Olig1), microtubule nucleation (C2CD3, ALMS1), cytoskeletal stabilization (CAMSAP1/2, Dystonin), axonal growth (DOCK7), axon guidance (Navigator 1/3), dendritic spine development (Formin1) or dendritic apparatus formation (Synaptopodin). [144–154] Similar proteins are also encountered in the presynaptic active zone (Piccolo, Bassoon), or the postsynaptic density (Shank1, LRFN2) of mature neurons and are important in the axonal transport of vesicles (Syntabulin, JIPs,), neurotransmitter release (Cacna1H, Mint2) as well as endocytosis (RUSC2). These extensive partnerships might explain why suppression of JNK activity was found to block differentiation of post-mitotic neuroblasts.[155] Ablation of total JNK activity is also embryonic lethal due to impaired neural tube closure.[37] Selective JNK1 (and to a lesser extent, JNK2 or JNK3) knockout mice suffer from a wide range of subtle brain abnormalities, including inappropriate cortical layer formation and tract defects.[156] On the other hand, JNKs also play a role in pathological axon degeneration after various insults. Knockout of JNK3 in mice results in enhanced ischaemic tolerance under experiments mimicking stroke. These kinases have already been suggested as attractive targets for the treatment of various neurodegenerative diseases

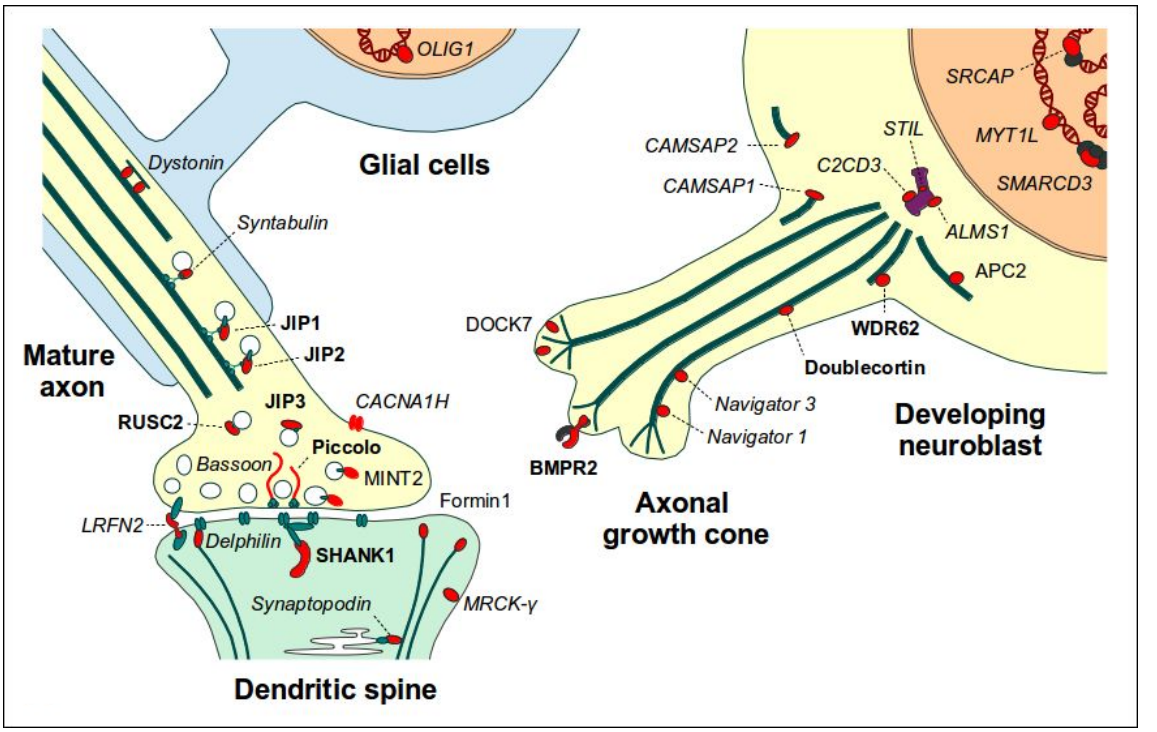


Figure 34: Proteins in developing neuroblasts (right) and mature synapses (right), already known to interact with JNK (bold letters), experimentally identified novel interactors (normal typesetting) and further, predicted JNK partners (italics).

The presence of insulin signaling pathway components on the lists can also explain many previous observations on the causative role of JNK in **type II diabetes**. This kinase forms part of the pathways over-activated by cytokines derived from adipose tissue. JNK1 knockout mice are also known to be resistant to type II diabetes induced by obesity.[157] The distribution of JNK-interacting proteins offers an elegant explanation for all these phenomena. Proteins bearing JIP1-type docking motifs are encountered on critical points of networks responsible for insulin signalling. These are the very same pathways that are also targeted by most anti-diabetic pharmaceuticals. Insulin secretagogues such as suphonylureas act directly on pancreatic beta cells. Insulin secretion is, however, controlled by MADD, one of the novel potential JNK partners, and a susceptability gene for type II diabetes.[158] Recombinant insulin analogs are used as a substitute to insulin itself , and act on the insulin receptor, whose critical downstream target, IRS1 (and IRS2) is another JNK partner.[159] Last but not least, insulin sensitivity can be increased by thiazolidinediones acting on the PPAR-γ receptor - but this also signals through a novel JNK partner protein, PGC1A. The negative regulatory role of JNK1 phosphorylation on IRS1 is well known, while the action of JNK1 on MADD is not. But the docking motif of PGC1A is mapped to the

same region, where phosphorylation was shown to induce degradation, thus it is not impossible that JNK also inhibits PGC1A, similarly to IRS1.[160]

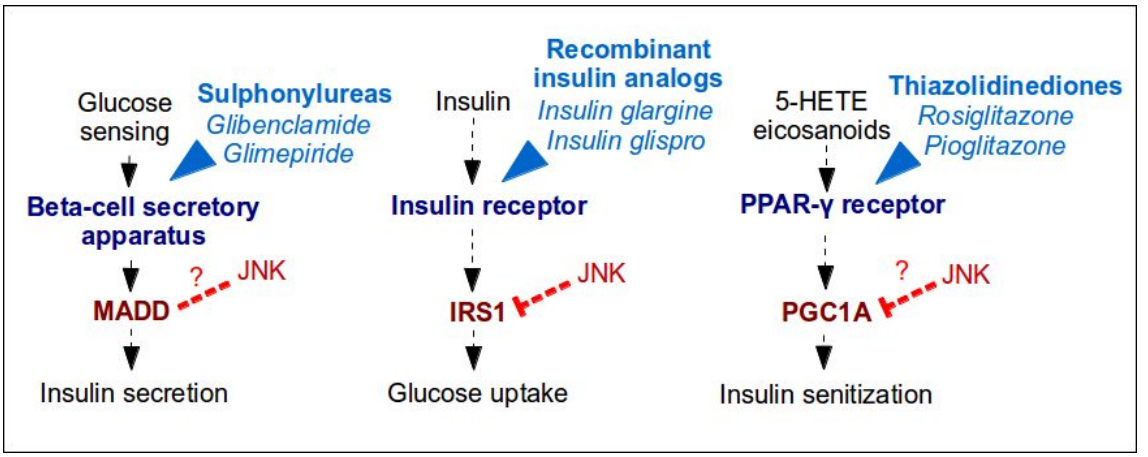


Figure 35: Proteins containing JIP1-type JNK docking motifs (red) control critical pathways in insulin signalling and type II diabetes, also targeted by anti-diabetic pharmaceuticals (blue)

The analysis of the best 100 hits for the **NFAT4** class yielded results comparable to the JIP1-type motifs, with some differences. However, members of the greater **MEF2A** class were markedly dissimilar from those of the JIP1 class. Here, the proportion of cytoskeletal proteins was minimal, while the fraction of transcription factors was considerably higher. Proteins involved in other gene-expression related functions, such as chromatin remodeling or histone methylation were also present in higher numbers. These differences were not surprising, since the latter motifs are predicted to mediate interaction towards p38α or ERK2, but not JNK1. When comparing distributions, the NFAT4 class appeared to lie in-between the two extremes represented by the JIP1-types (mostly cytoplasmatic targets) and greater MEF2A types (emphasizing nuclear actions). The similarity of NFAT4-type motif containing proteins to JIP1 type bearing ones is easy to understand: both primarily interact with JNK1. In certain protein families, one can discover closely related pairs in which one protein contains a JIP1-type docking motif, and the other contains a likely independently evolved, NFAT4-type docking motif. On the other hand, the NFAT4-type motifs are structurally compatible with MEF2A-types (unlike JIP1-types), thus some of the predicted best binders are shared between the latter two lists. Our dot-blot experiments indeed corroborated that the overlapping motif definitions result in a naturally overlapping set of interactors for JNK1 and p38α (see supplementary information).

Experiments with greater MEF2A-type motif bearing proteins provided another interesting observation. The occurrance of such motifs in the **AMPK pathway** implies that this system connects to the p38 and/or the ERK1/2 pathways on very specific points, in specialized tissues. The AMPK pathway lies at the intersection of nucleotide metabolism and energy homeostasis.[161] During periods of excessive energy use and ATP depletion, the nucleoside-diphosphate-kinase enzyme helps to produce ATP at the expense of ADP. Adenosine-monophosphate (AMP) is also released, setting a chain of regulatory events into motion. AMP-activated protein kinase (AMPK) is turned on, phosphorylating a variety of metabolic enzymes and transcription factors - in order to mobilize reserves and reduce energy consumption. Kinase suppressor of RAS (KSR) proteins are also substrates of AMPK. The two KSR proteins play very different roles: KSR1 is responsible for the control of growth factor pathways. On the other hand, KSR2 is expressed predominantly in the brain, and ample evidence suggests that it controls energy expenditure and feeding behaviour.[162] Excess AMP is partly eliminated by adenosine monophosphate deaminase (AMPD) enzymes - turning AMP into inosine-monophosphate (IMP) and ammonia. According to our experiments, p38α and ERK2 specifically interacts with the cardiac AMPK enzyme (bearing the AAKG2 regulatory subunit).[163] However, the AMPK enzyme found in most other tissues lacks this MAPK binding motif. Through similar motifs, p38α also has the potential to interact with the skeletal muscle-specific AMPD1 or the cardiomyocyte- and erythrocyte-specific AMPD3 enzyme (but not with the broadly expressed AMPD2). ERK2 - on the other hand - has the ability to tightly regulate KSR2, a protein with a central nervous system focused expression pattern according to gene expression databases.[164] Interaction with tissue-specific protein isoforms may thus cause highly cell-type specific regulation by MAPKs.

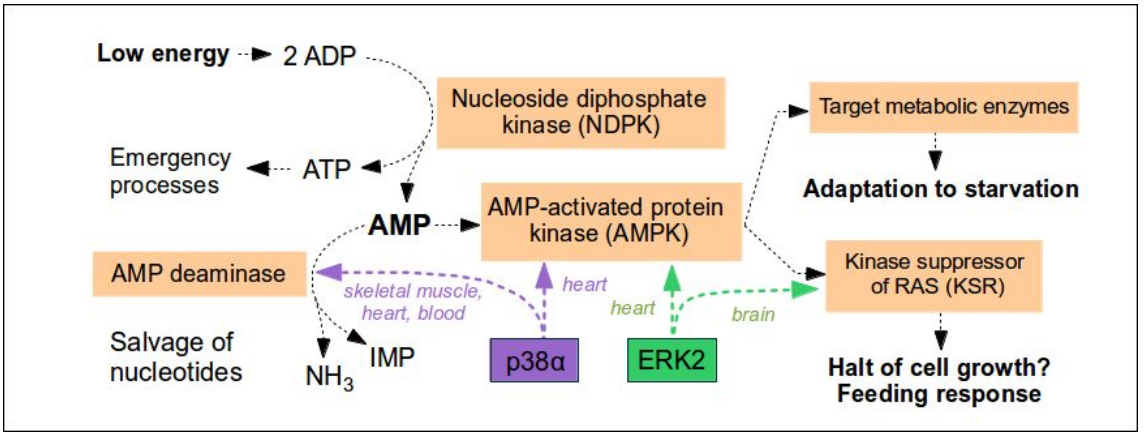


Figure 36: Schematics of the AMPK pathway and its proposed, tissue-specific regulation by p38α and ERK2 (at either identical or different points of the network).

Evolutionary analysis of hits

MAPK pathways are found in almost every eukaryotic organism and the three-tiered kinase cascade architecture of the MAPK module core is well-conserved from yeast to human. Therefore one would naturally expect the downstream targets of these pathways to be **conserved** as well. However, our results suggest the opposite. We calculated several conservation parameters for this purpose, including conservation of core motif resides, the same versus the conservation of flanks, etc. This included calculation of the maximum traceable distance (MTD): the value of evolutionary distance to the most divergent organism the motif is found in (versus human). These automated conservation parameters were designed and calculated by Olga Kalinina (MPI, Saarbrücken), using pre-computed alignments from the EggNOG and InParanoid databases.[165,166] Unfortunately, due to low reliability of automated alignments with sequences from non-vertebrate organisms, at the end we limited ourselves to vertebrate sequences only (the most distant organism used was zebrafish, *Danio rerio*).

Although linear motif conservation is often used for prediction purposes, the same approach did not work when dealing with D-motifs. We failed to detect correlation between FoldX (predicted binding energy) and any of the evolutionary conservation scores. The maximum traceable distance (MTD) of a motif in evolutionarily related species could be calculated from the EggNOG alignments. Here we also noted that most of the motifs were traceable to vertebrates only. A more thorough search, using p-Blast searches in the UniProt database revealed that some motifs are actually more ancient than what EggNOG data would suggest. Still, a high number of experimentally validated motifs were found to be relatively recent evolutionary inventions. After mapping the most distant organisms in which the motif in question is already present, we were able to compile an evolutionary histogram of MAPK docking motif emergence (see figure).

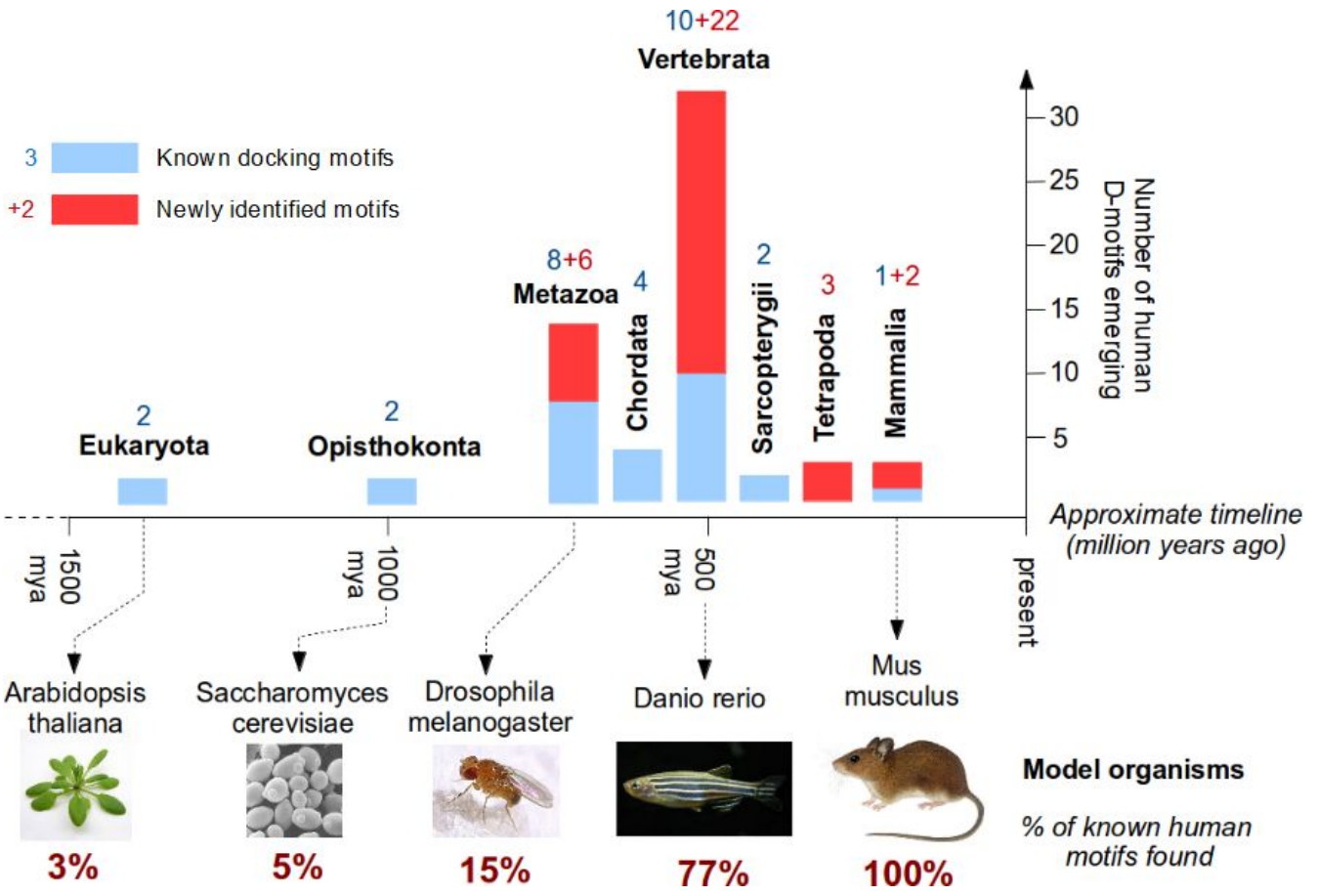


Figure 37: Histogram summarizing the statistics of human D-motif emergence. Blue columns and numbers refer to protein families with a previously-known motif instance, while red columns and numbers represent proteins with newly identified motifs (validated at least at a fragment level).

Despite the fact that MAPK pathways are an eukaryotic common heritage, very few human docking motifs had an ancestry among unicellular organisms. This was only true for the MAPK kinases (MAP2Ks) or MAPK activated kinases (MAPKAPKs) and a few truly ancient substrates, like MEF2/MADS-box proteins. Only in multicellular animals (Metazoa) did docking motifs become detectable on a variety of phosphatases and MAP3Ks as well as on the core set of mammalian substrates (ELKs, ATFs, JUNs, etc.). However some of these motifs were difficult to find as they were subsequently lost on several lineages, especially in arthropods (e.g. the tendency of motif loss was obvious in *D. melanogaster* sequences) . The diversification of docking motifs continued in chordates: But it is the early vertebrate evolution, where a major re-wiring and expansion of MAPK partnerships occurred. Over 50% of the motifs identified in our experiments evolved at this period. After the development of bony fishes, motif emergence events became less common, but did not stop completely: New motifs kept appearing in lobe-finned fishes (Sarcoperygii), in terrestrial vertebrates

(Tetrapoda) and even in mammals. Comparison of the known and predicted motifs from the best 100 hits for JIP1-type motifs suggests that there are many more recently-evolved motifs in mammals, just waiting to be discovered (see figure 38, panel A). These findings are well in-line with recent results on yeast calcineurin interactomes. Yeast phosphatase docking motifs were found to evolve fast and their distribution was highly divergent between related species.[167]

A further proof for the late evolution of MAPK partnerships is found when comparing **paralogs**. These latter are closely related copies of the same ancestral gene that often preserved linear motifs from before their split. Most vertebrate proteins come in groups of 2, 3 or 4 closely related paralogs due to twin genome duplications - and subsequent gene loss - at the dawn of vertebrate evolution.[168] Interestingly, most of the better-known MAPK target proteins possess a D-motif in more than one vertebrate paralogs. However, the same is not true for the majority of novel partner proteins. Comparison of vertebrate proteins with those from earlier-branching genomes also helped us to determine if a motif developed after the gene duplications or did not. Our statistics suggest that the presence of more than one paralog with the same motif is predictive of an ancestral pre-vertebrate motif. (Over 50% of such protein families have non-vertebrate members with the motif already in place.) In this case, motif loss appeared to be the dominant mechanism to create differences between vertebrate paralogs. Only a very few new motifs emerged in-between the two whole genome duplication events, suggesting that this evolutionary stage was short-lived. On the other hand, where only a single paralog contained the motif, this motif was overwhelmingly a new invention after the twin duplications - and not a result of an ancient motif being lost. The latter one was the single most common scenario, especially with newly found D-motifs. Many of these novel MAPK recruiting motifs are suspected to provide a paralog (or even isoform) specific regulation. This could offer unique roles to otherwise highly similar human proteins.

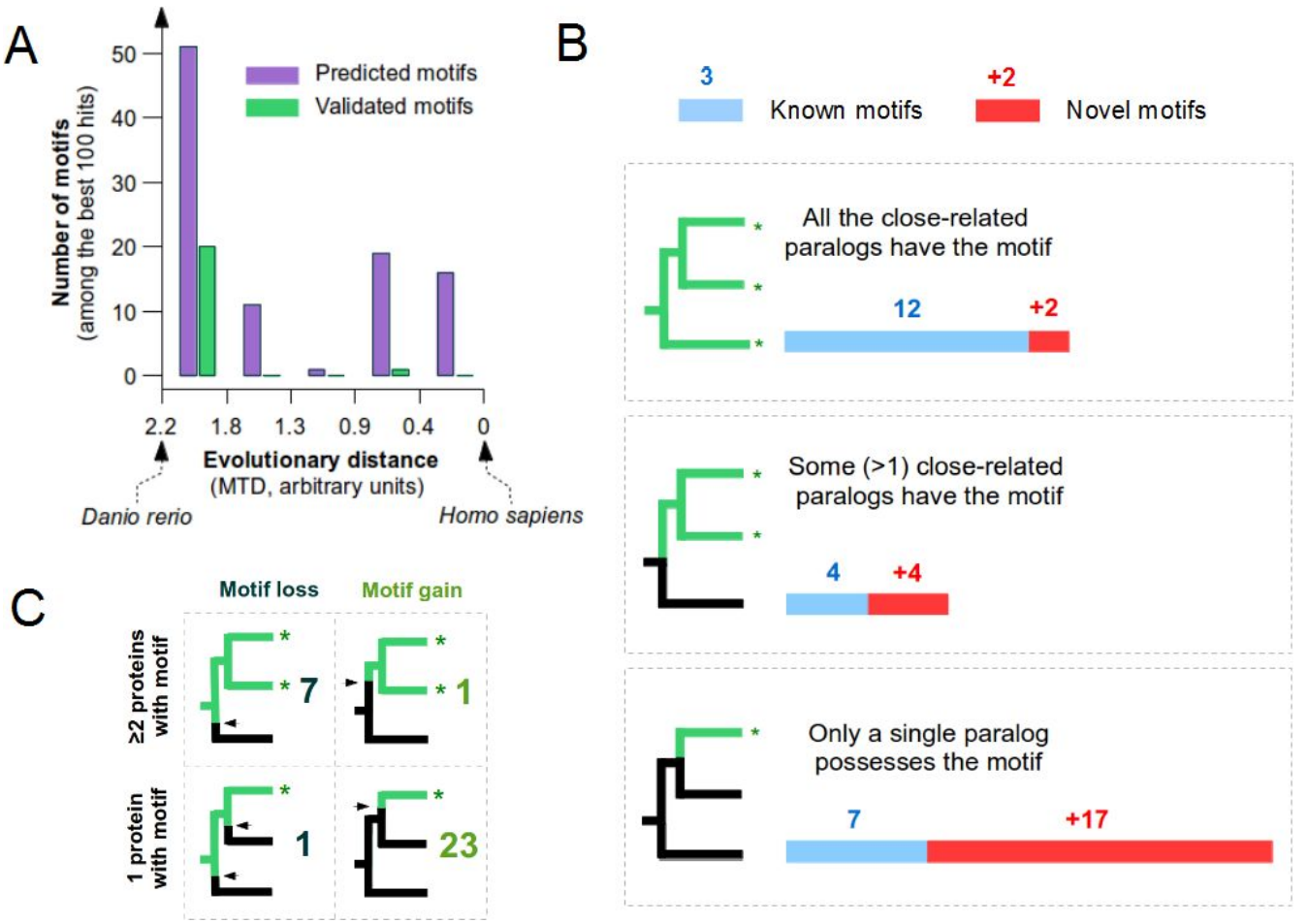


Figure 38: The emergence distribution of known and predicted best 100 motifs in the JIP1 class suggests that recently-evolved motifs are still under-sampled (A). Human D-motifs may be found in some or all closely-related paralogs of the protein; although newly-dscovered motifs strongly suggest that stand-alone motifs are the most common (B). Motifs that are found in more than one paralog are usually a pre-vertebrate heritage, while stand-alone motifs typically evolved after the twin genome duplications, hinting at rampant D-motif emergence in early vertebrates (C)

Having obtained a sufficient number of experimentally verified examples, we could also test some theories on the **evolution of linear motifs**. The motifs we validated (at least at a fragment level) could be classified based on their predicted origins. Not surprisingly, the most common way of motif emergence appeared to be random mutations in an already-existing disordered segment. This could be illustrated with a known interactor, the Smoothelin-like protein 2 (SMTL2).[169] Here, gradual sequence changes in terrestrial vertebrates led to the creation in the motif, which is restricted to placental mammals (Eutheria). There were also several examples for creation from scratch (i.e. non-coding DNA). This could mean either translational start shift (translating an earlier non-translated 5' UTR) or splicing site shift (leading to exonization of intronic sequences). While the N-terminal expansion of the protein is seen in MCL1 (the motif-bearing segment has no counterpart in Bcl2 or in any other related

protein), another newly identified partner, KSR2 serves as an example for splicing site rearrangements. Here, the paralog KSR1 retains the ancestral intron-exon boundaries, which appear to have shifted in KSR2. We could even find examples for proteins where the mechanism was still active: the motif can (in an isoform-specific way) be included or excluded due to alternative splicing or initiation. This is the case with the PDE4 genes, where most paralogs (PDE4A, PDE4B, PDE4D) still retain an ancestral, alternative exon containing a JIP1-type motif (see the supplementary for details).

Interestingly, linear motifs can also transmute into each other: some examples in the dataset show potential switching between different MAPK docking motif classes. As a result, distant organisms may show different motif types at the same spot: i.e. the JIP1-type motif we identified in MKP5 corresponds to an NFAT4-type motif in distant organisms (as in arthropods). In contrast, CCSER1 has an NFAT4-type motif in humans, but a JIP1-type one in zebrafish. The motif in ELK1 is of the JIP1-type in humans, but Far1-type in protostomes. Also, the atypical (incomplete) motif of TAB1 is DCC-like in most non-vertebrate organisms, unlike the human which is MEF2A-like. In some cases, a “horizontal motif transfer” (i.e. recombination between unrelated genes) may have complemented the de novo emergence of motifs. This was likely the case for AAKG2, where the similarity of the entire N-terminal region of AAKG2 (which is unique and sets it apart from AAKG1) with the C-termini of MEF2A or MEF2C was surprising. The disordered segments flanking the motif also align well. This cannot be explained by convergent evolution alone, since those segments are not subject to the same selection. The creation of a new linear motif from the unfolded remnants of earlier structured domains was yet another intriguing possibility, although it was definitely rare. For the WDR62/MABP1 family, the duplication of WD40 repeats and their subsequent degeneration were the most likely source of the NFAT4-type D-motif (see supplementary).

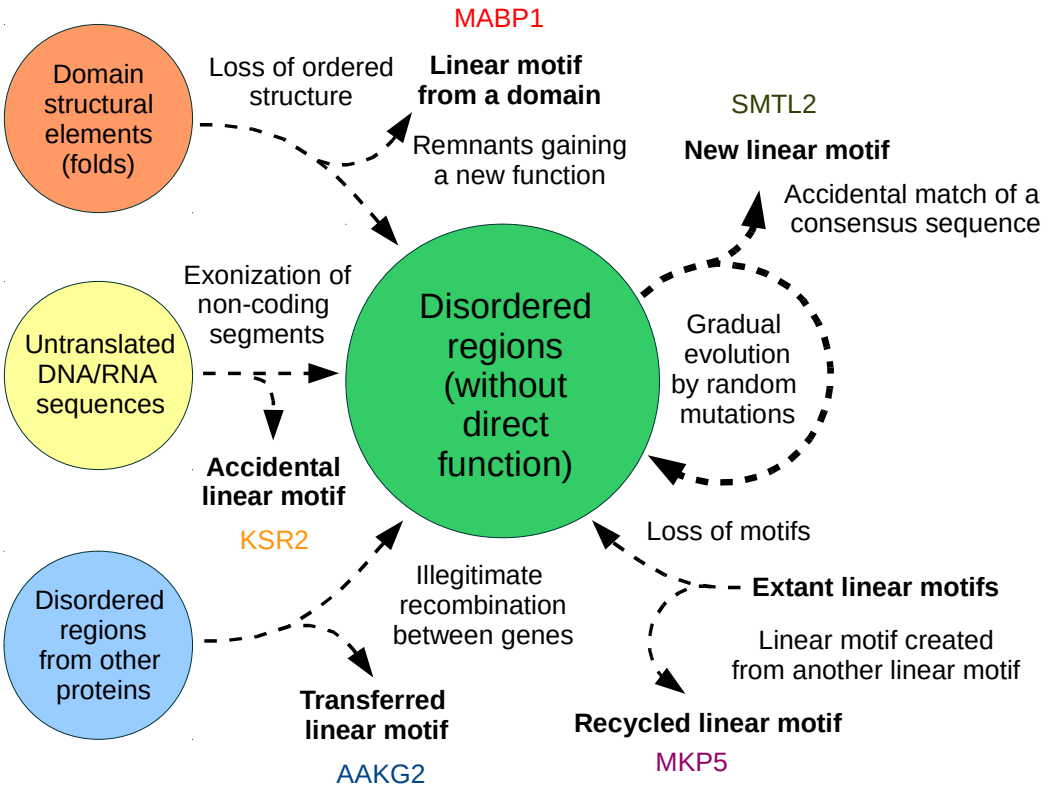


Figure 39: Diverse evolutionary origins of MAPK docking motifs, with examples picked from newly-discovered D-motifs. Emergence through genetic drift in non-conserved, intrinsically disordered segments appears to be the most common mechanism (thick dotted lines)

Functional aspects of docking motif evolution

The typical purpose of docking motifs is to enable phosphorylation of recruited substrates. Due to the lack of strict spatial constraints between D-motifs and phosphorylation sites, such roles can only be interpreted in the broader context of a protein. Unfortunately, most **phosphorylation target sites** controlled by the novel docking motifs remain elusive. Yet in some rare cases, such sites have either been discovered beforehand or could be inferred based on spatial proximity to the D-motif and their co-evolution.[169–171] Analysis of those examples gives a dramatic insight on how docking motifs emerged in relation to their target sites. In particular, both motif loss and gain is expected to have a profound effect on target sites: potentially endowing the protein with a new regulation. This is well illustrated by three

cases: the Nuclear factor of Activated T-cells (NFAT) family, with a de novo motif created in NFAT4 (adding on to a pre-existing target site), the Myocyte Enhancer Factor 2 (MEF2) family, displaying motif loss to a varying degree (with the concomitant loss of target sites), and the Grb2-Associated Binder (GAB) family, in which both events took place (see the figure below).

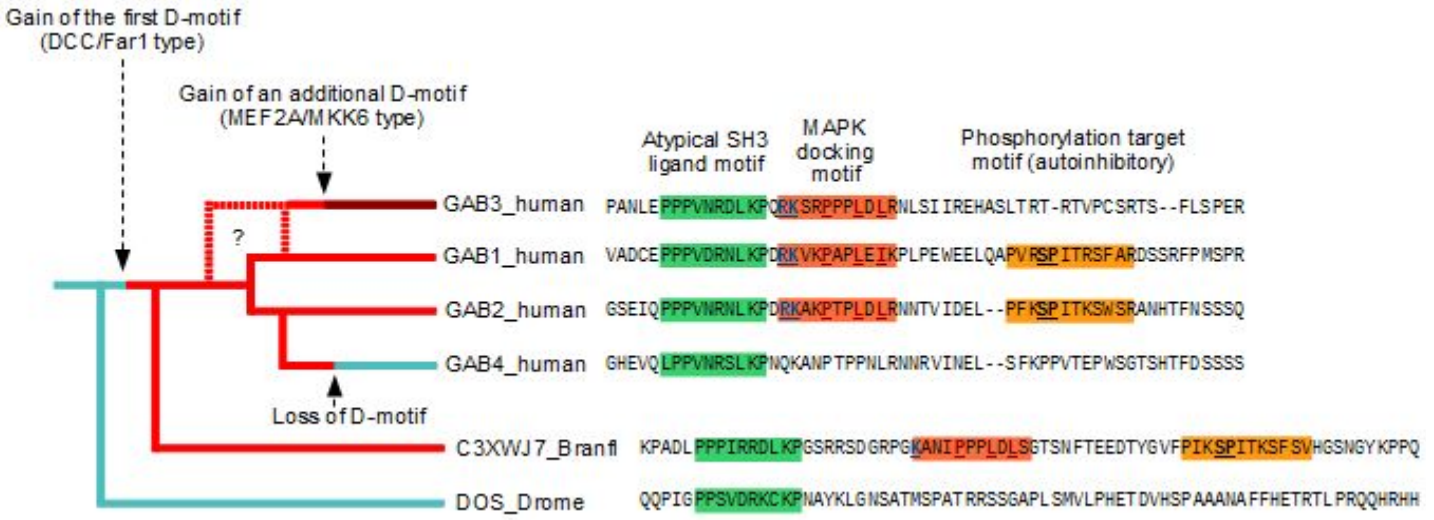


Figure 40: Suggested evolution of the MAPK-dependent regulatory modules in Gab proteins. The atypical SH3 ligand motif - binding to Grb2 (green) - is always present, while the D-motif (red) and the autoinhibitory motif regulating lipid binding of the PH domain (orange) are probably more recent inventions, found only in chordates. (The additional D-motif of Gab3 and its putative target sites lie outside the region shown here. Branfl= Branchiostoma floridae, Drome = Drosophila melanogaster)

In the case of Gab1, the ERK2 target motif lies in close proximity to our newly-identified docking site, and its modification controls an intramolecular interaction: When phosphorylated, the target motif is no longer able to act as an autoinhibitory element, releasing the N-terminal PH (pleckstrin homology) domain to allow phosphatidyl-inositide binding and membrane association.[170,171]

Unfortunately, it is usually not straightforward to assess the effects of phosphorylation on a particular protein at a particular site. Even if phosphorylation events happen directly in the vicinity of known interaction motifs, it may or may not have an influence on that binding event. To test the possibility whether phosphorylation of the C-terminal disordered segment of DOCK5 has an effect on its binding to CrkII (a known interactor of this region, using its N-terminal SH3 domain to bind any of the three SH3-ligand motifs located here)[172] , I have cloned both proteins and tested the possibility in a GST

pulldown assay: The C-terminus of DOCK5 was pre-phosphorylated by incubating the glutathione-sepharose beads coated with DOCK5-CT in kinase buffer and activated MAPK for 2 hour (this can also be seen on the bandwith shift on SDS-PAGE). However, the phosphorylation of DOCK5 had no effect on its binding to CrkII (see figure).

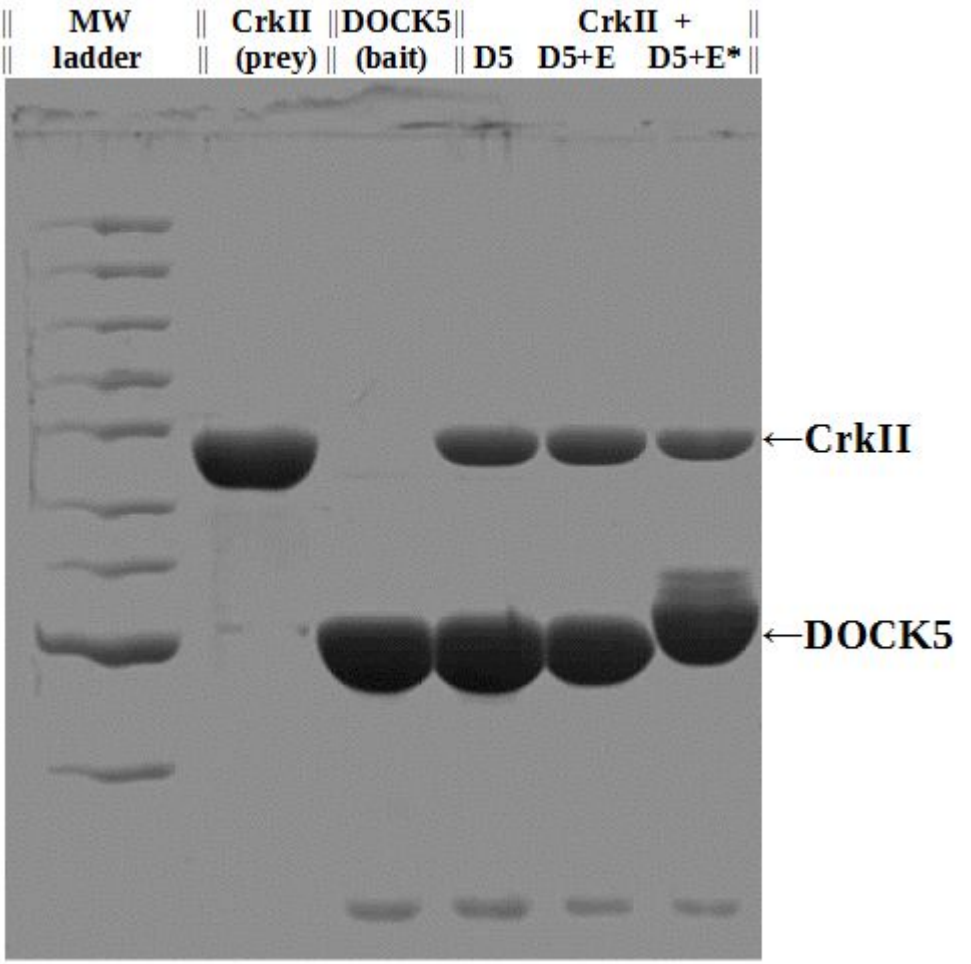


Figure 41: Pull-down experiment showing the lack of phosphorylation-dependent effects on the interaction of the flexible C-terminus of DOCK5 with CrkII. The DOCK5 (D5) constructs were immobilized on GST beads and treated differently, either subjected to a direct pull-down, or pre-incubated with enzyme (E) for 2 hours in kinase buffer without or with ATP (*) prior to pull-down. despite the band shift indicating phosphorylation (and even partial hyperphosphorylation)

Our studies also support the notion that there are a number of proteins with **multiple MAPK docking elements**. Apart from the case where these elements interact with different MAPKs (i.e. in the case of MKP5, where a rhodanese domain can bind p38 and a JIP1-type linear motif interacts with JNK; or Gab3, with separate motifs to recruit ERK1/2 or p38), the purpose of multiple D-motifs is unclear.

Especially because they tend to bind to the same surface and thus compete with each other for MAPK binding. Such domains or motifs are often not simple duplicates of each other and emerged independently during evolution. This happened in the case of BMPR2, where the first JIP1-type motif is found in almost all multicellular animals, but the second one is restricted to vertebrates, and the ATF2/CREB family of transcription factors, where the N-terminal motif embedded in a Zn-finger is the primary docking element in all metazoans, but several vertebrate paralogs have an additional JIP1-type motif with an unclear role.[173,174] As one MAPK molecule can only accommodate one motif at a time, it is probable that multiple docking motifs would allow several, mutually largely exclusive complexes - each with unique spatial orientation. As in the case of MKK7 which activates JNK1, the precise orientation of the MAPK versus the partner protein might have important implications on phosphorylating specific target sites.[175]

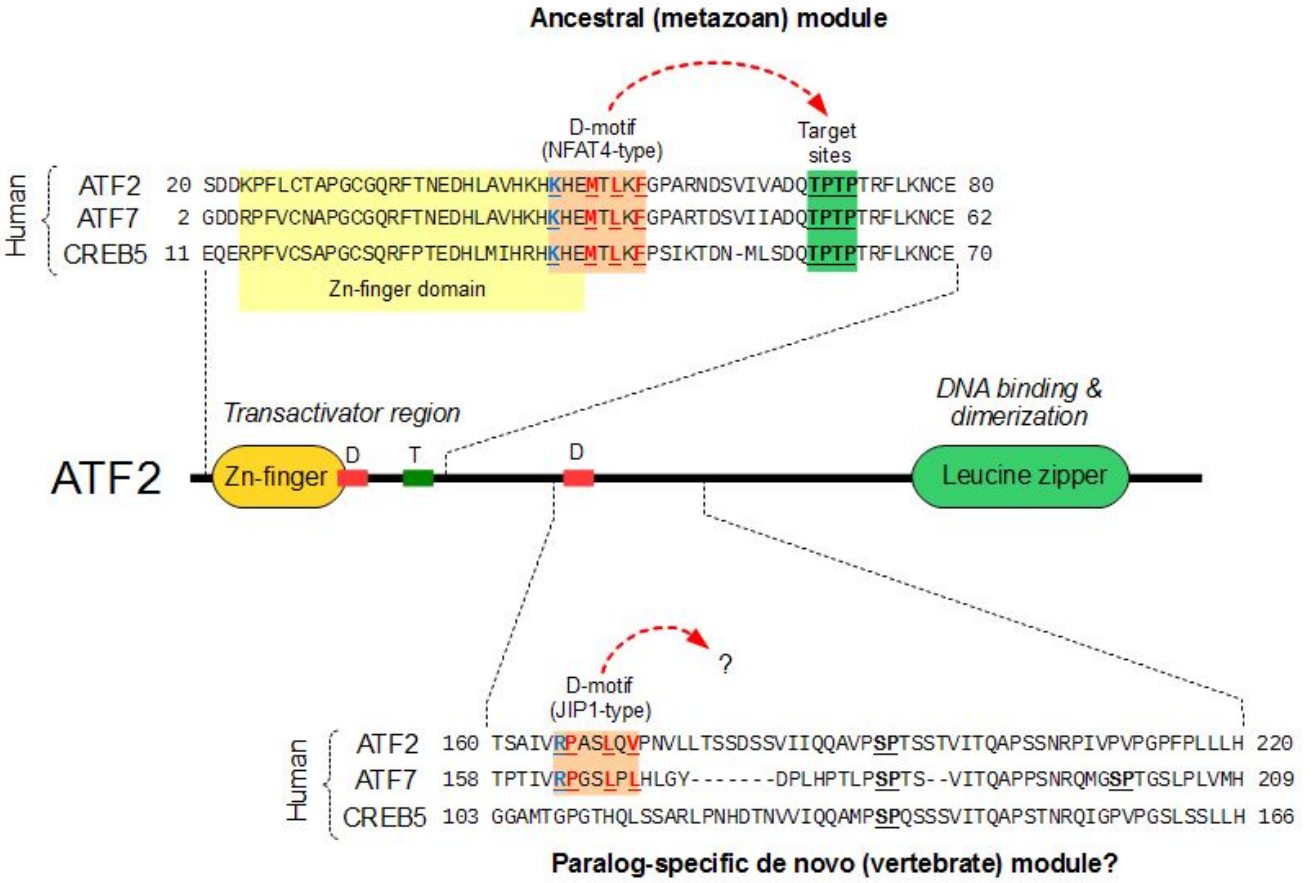


Figure 42: The ATF2 family of transcription factors harbors a pair of ancient MAPK-docking and phosphorylation target sites. Unlike non-vertebrate proteins and the paralog CREB5, ATF2 and the closely related ATF7 shows a de novo emergence of a JIP1-type docking motif in a separate part of the protein, neighbouring several SP/TP sites.

Exploring atypical motifs and non-motifs

Finally, I also studied a number of intriguing examples that cannot be predicted by our systematic searches. Although there are many, more-or-less well-known cases of D-site binding proteins , the absolute majority of them utilize canonical D-motifs. The remaining few outliers - on the other hand - use alternative solutions to interact with the same docking site of MAPKs. In some cases, a D-motif (or at least parts of it) can be detected, but it cannot mediate binding alone. In this case, a common biological solution is to use extra motifs or domains to complement a defective docking motif. Unfortunately, these cases cannot be truely predicted by current methods. In other cases and other proteins, the D-site interacting elements are either partially or completely folded, thus they are no linear motifs at all, but true "D-domains". In order to study a few already-known “atypical” examples, I also performed a series of simple experiments. Three systems will be shown here: each displaying a different type of deviation from the canonical D-motifs.

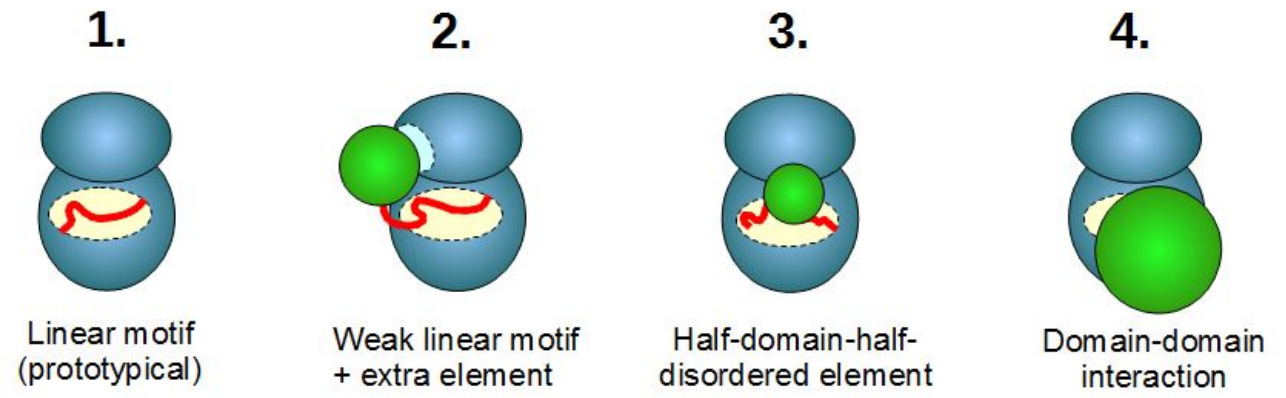
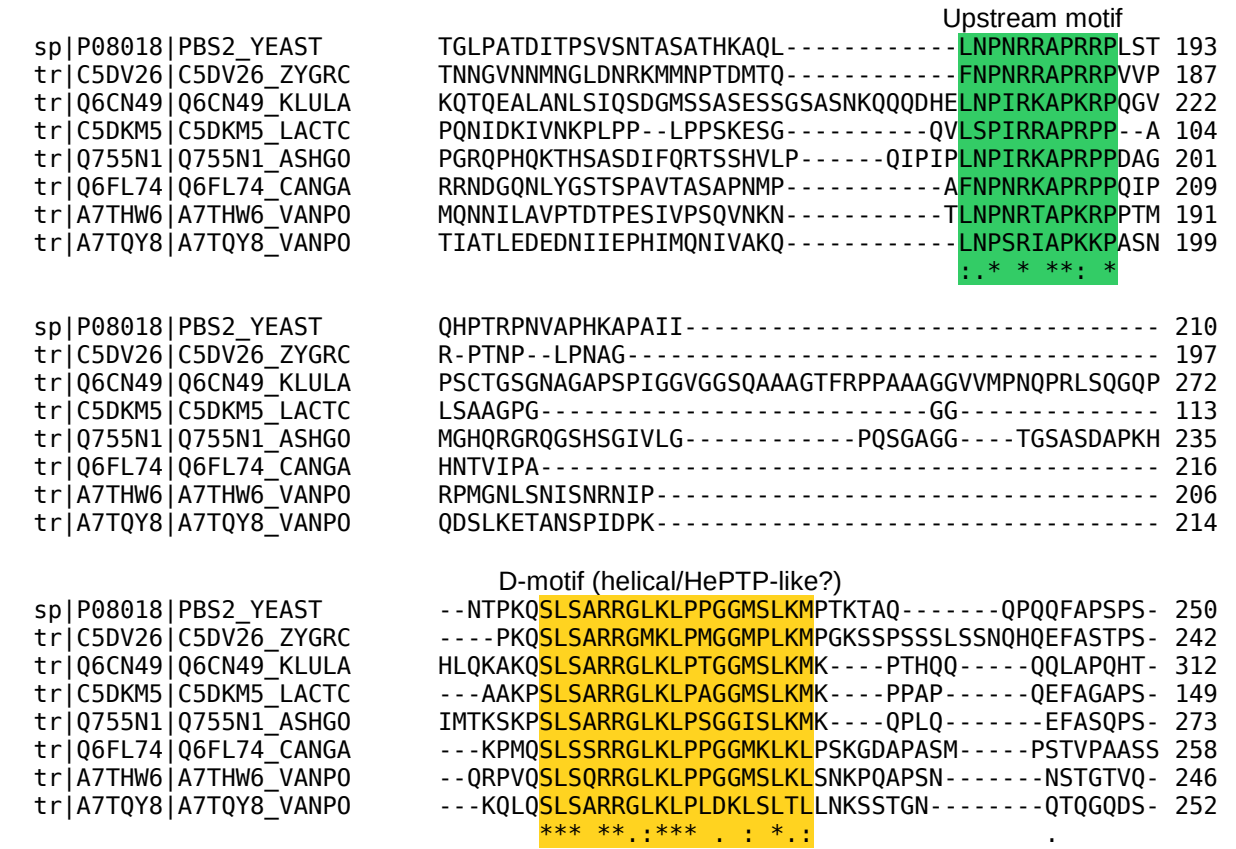


Figure 43: Prototypical D-motifs (1) and different types of deviation from canonical linear motifs (2-4). Some linear motifs are inherently defective and synergize with extra domains in order to be able to interact with their partner at a physiological strength (2). Other interaction elements appear to be a "cross" between a domain and a motif, encompassing both folded and intrinsically disordered segments binding to the same site in a contigous manner (3). Finally, a number of folded domains can also evolve specialized surfaces to associate with a binding site, usually filled out by a liner motif in all other cases (4).

1) **The Hog1-Pbs2 interaction**

In the yeast *Saccharomyces cerevisiae*, the Hog1 pathway is responsible for adaptation to hyperosmotic conditions as well as to respond to a number of other stress stimuli. As already mentioned in the introduction, the middle tier of this pathway is provided by the MAP2K kinase Pbs2. Although this fungal pathway is homologous to mammalian p38 and JNK signalling, the interaction between this MAPK-MAP2K pair displays some peculiar characteristics. Despite the fact that Pbs2 possesses a well-conserved docking motif on its disordered N-terminus, it is insufficient to mediate binding of its substrate, Hog1. Previous studies have established that a broader region surrounding the motif is both necessary and sufficient for this enzyme-substrate interaction. This “Hog1 binding domain” or HBD region is however, not a true domain: Predictions by IUPRED suggest that the entire region is intrinsically disordered. Evolutionary conservation analyses similarly point to the existence of two short linear motifs in the region, and the lack of conserved domains. One of the motifs is a more-or-less typical D-motif (either greater MEF2A or greater HePTP-type). The other one is a short proline-rich segment, known to function as an SH3-domain ligand.

To check if both motifs are required to bind Hog1 directly, I decided to further refine mapping of the HBD region. To this end, a GST-pull down experiment was performed with three, bacterially expressed constructs. One of them was the previously-established HBD region. The second one consisted of the minimal segment encompassing both motifs. The third one only contained the C-terminally situated D-motif with its broader surroundings (see figure on the next page).



HBD-full construct:

QRMSSQVVQASSKSTLKNVLDNQETQNITDVNINIDTTKITATTIGVNTGLPATDITPSVSNTASATHKAQLLNPNRRAPRRPLSTQHPTRP
NVAPHKAPAIINTPKQSLSARRAVKLPPGGMSLKMPTKTAQQPQQF

HBD-mid construct:

AQLLNPNRRAPRRPLSTQHPTRPNVAPHKAPAIINTPKQSLSARRAVKLPPGGMSLKMPTKTA

HBD-pep construct:

QSLSARRGLKLPPGGMSLKMP

Figure 44: Alignment of Pbs2 kinases from related yeast species, showing two well-conserved short motifs in the so-called HBD (Hog1-binding "domain") region. [YEAST = Saccharomyces cerevisiae, ZYGRC = Zygosaccharomyces rouxii, KLULA = Kluyveromyces lactis, LACTC = Lachancea thermotolerans, ASHGO = Ashbya gossypii, CANGA = Candida glabrata, VANPO = Vanderwaltozyma polyspora]. Below, the amino acid sequence of the three S. cerevisiae constructs used in the GST-pull-down experiments are shown.

The pull-down experiment against recombinantly expressed and purified Hog1 (from SF-9 cells, with the help of Anita Alexa) clearly shows that the second fragment provides a relatively strong interaction similar to the one seen with the complete HBD region. On the other hand, the D-motif alone does not bind to Hog1 with a comparable affinity. (The reverse experiment, using on the proline-rich segment alone was not done; as it is already established that the D-motif of Pbs2 is required for Hog1 binding). Thus it is likely that here, two different linear motifs act synergistically to bind the yeast Hog1 MAPK.

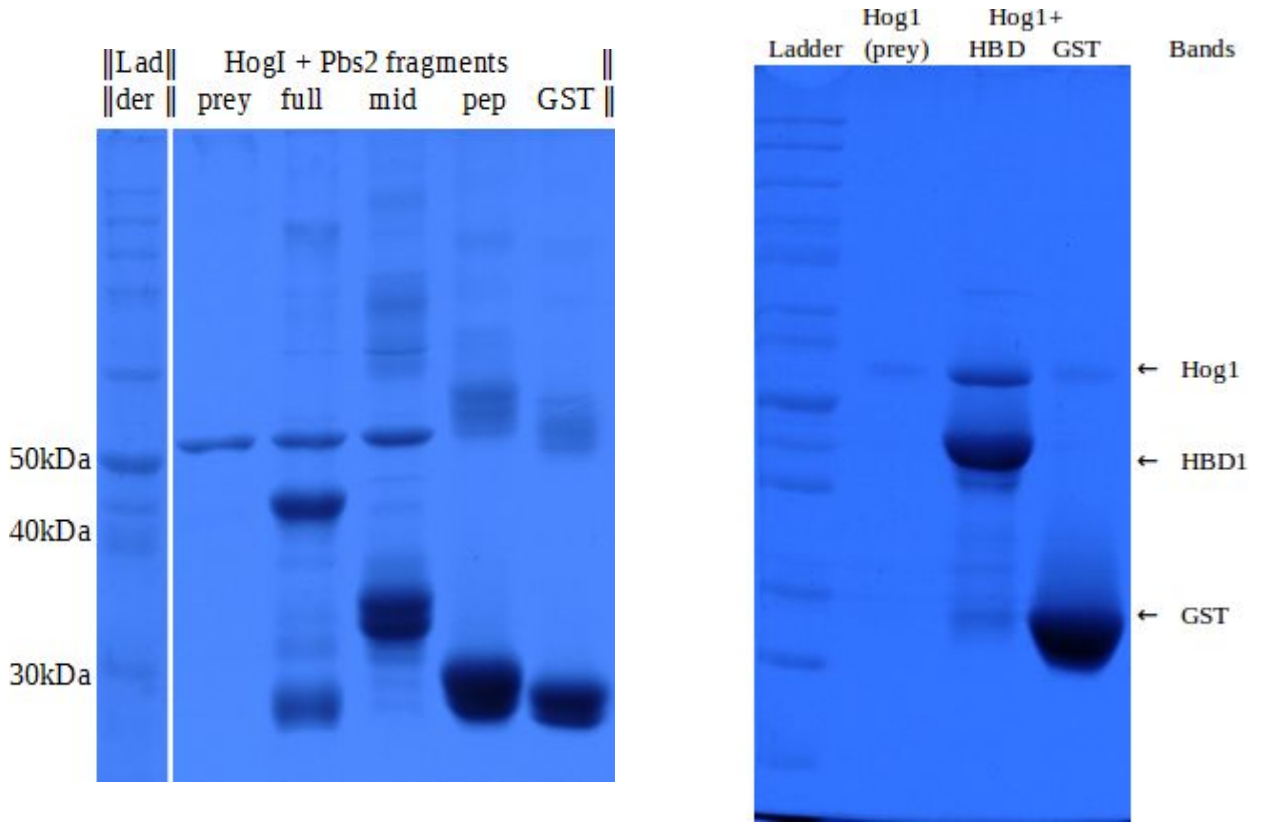


Figure 45: GST pull-down experiment showing that both conserved motifs in the Hog1-associating HBD region of Pbs2 are required for appropriately strong Hog1 binding (Coomassie-stained SDS-PAGE, "full" construct = HBD region = Pbs2[108-245], "mid" construct = Pbs2[177-239], "pep" construct = Pbs2[215-235]).

2) **Semi-rigid motif of ATF2**

Activating transcription factor 2 or ATF2 is one of the oldest known mammalian MAPK (and more specifically, JNK) substrates. It is also known that phosphorylation of its N-terminal transactvation motif requires classical D-site mediated docking. However, the D-motif in this case partially overlaps with a Zn-finger domain. The NMR structure of the N-terminus of ATF2 clearly shows that the C-terminal part of the motif is genuinely disordered in solution (as a linear motif should be).[176] On the other hand, its N-terminus forms part of a rigid alpha-helix, tightly held by the Zn-binding residues located nearby. Although the ATF2 was previously demonstrated and generally accepted as possessing a D-motif, its structure was never examined in detail. In fact, the way how such an aberrant "half-motif - half domain" element binds to its target protein is still somewhat unclear. An X-ray structure allegedly showing JNK3 bound to the D-motif of ATF2 has recently been published, but it is rife with controversies.[115] The peptide that is thought to represent the D-motif is bound to the surface of JNK3 in a way that is incompatible with Zn-ion coordination. The authors also noted that their peptide bound with rather low affinity (Kd>10μM, as determined by isotermic calorimetric titrations).

To test if their approach was appropriate, we decided to re-map the docking element on the N-terminus of ATF2. In a preliminary pull-down experiment, I have shown that the N-terminus of ATF2 strongly interacts with JNK1, but only weakly with ERK2 or p38α. However, in the same experiments, use of the additive EDTA (a strong metal ion chelator, in 20mM concentration) caused a strong reduction in the ATF2-JNK1 interaction. As similar chelators are routinely used to denature Zn-fingers, we may conclude that in this case, the Zn-finger contributes to the MAPK binding. But in order to understand its extent of importance, a panel of 8 GST-fusion constructs were produced, encompassing different fragments of ATF2. The results of this fine-scale mapping experiment undoubtedly shows that an intact Zn-finger is required for "wild-type" affinity (see the figures below).

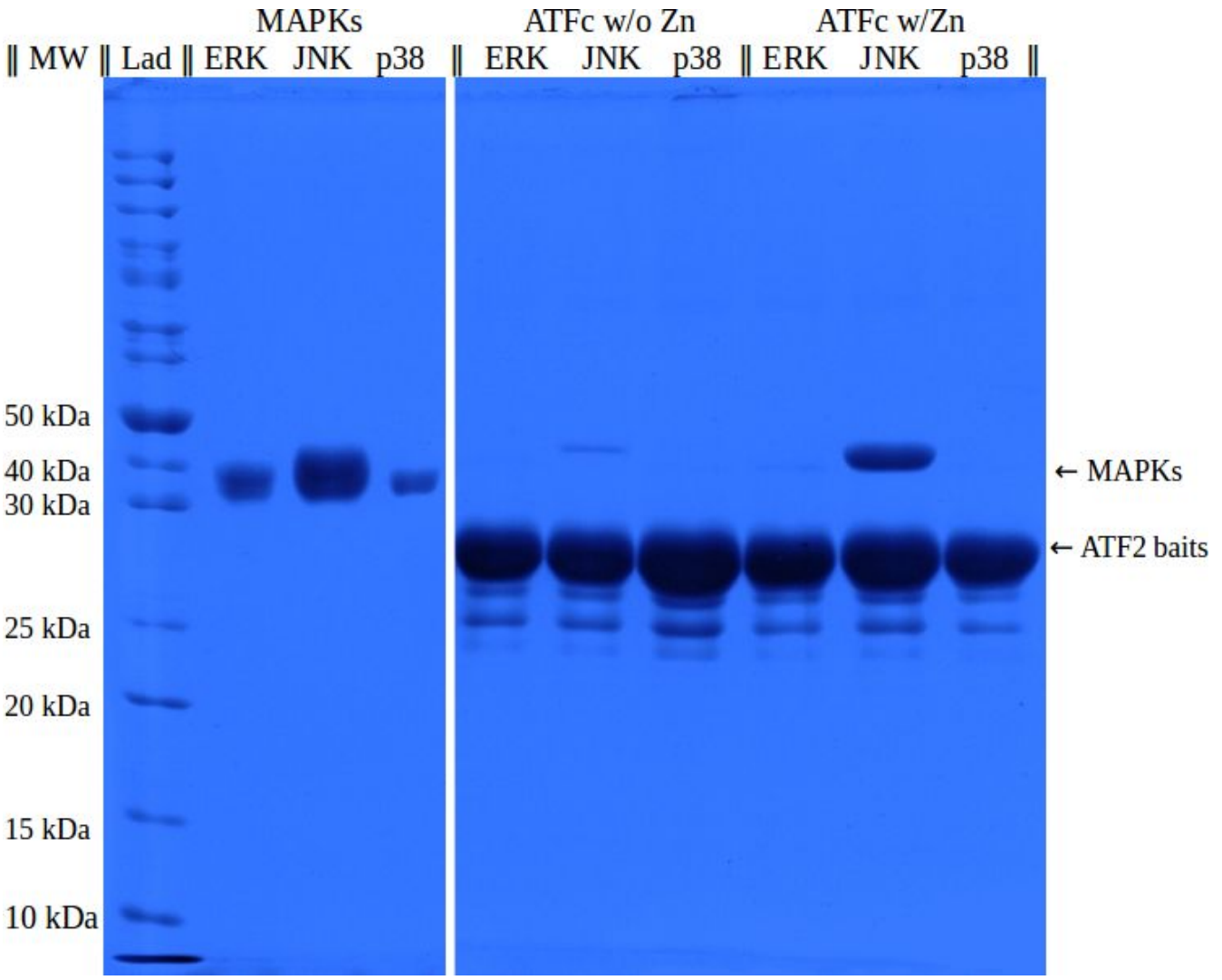
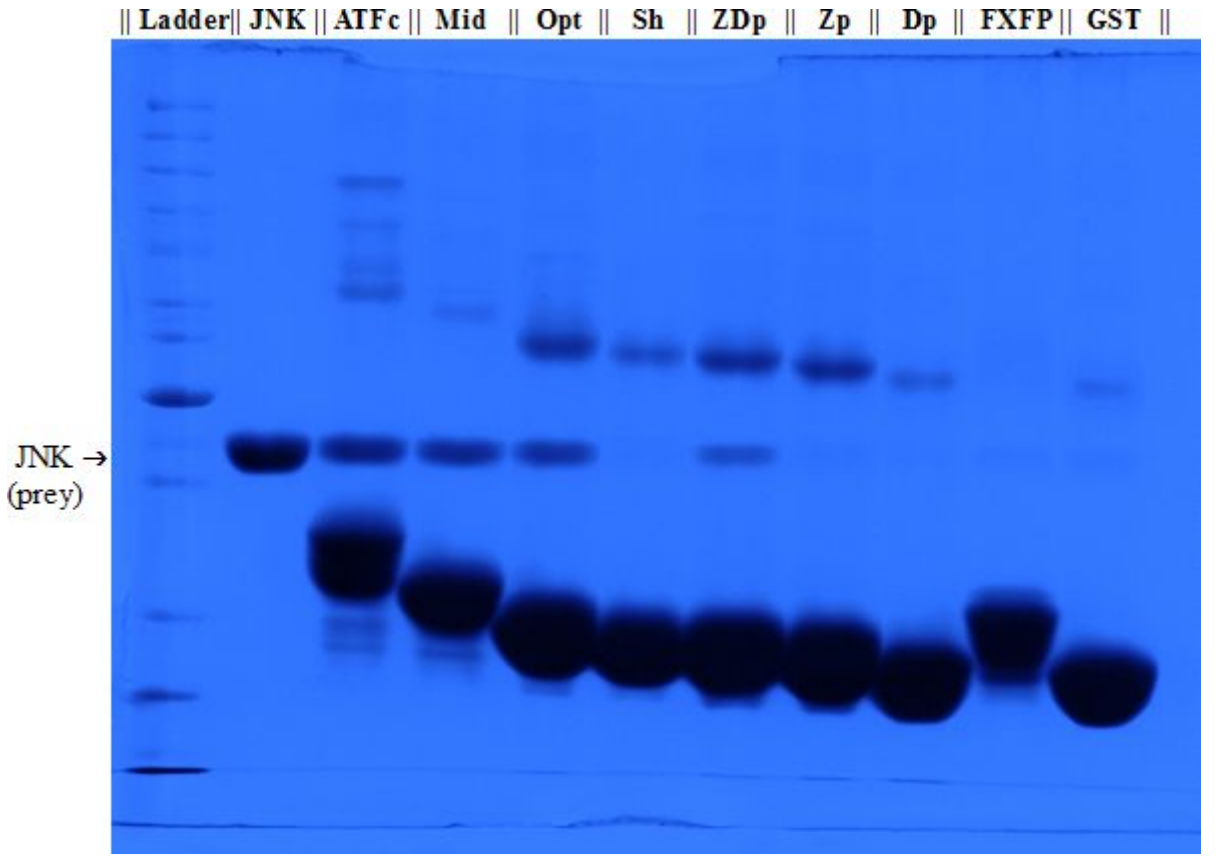


Figure 46: GST pull-down experiment with the complete transactivation region of ATF2 (ATF2c = ATF2[19-99]) against ERK2, JNK1 and p38α, in a buffer containing the strong chelator EDTA (middle lanes) or under non-chelating conditions (right lanes)




```
ATFc: MSDDKPFLCTAPGCGQRFTNEDHLAVHKHKHEMTLKFGPARNDSVIVADQTPTPTRFLKNCEEVGLFNELASPFENEFKKAS
Mid:  MSDDKPFLCTAPGCGQRFTNEDHLAVHKHKHEMTLKFGPARNDSVIVADQTPTPTRFLKN----------------------
Opt:  MSDDKPFLCTAPGCGQRFTNEDHLAVHKHKHEMTLKFGPA------------------------------------------
Sh:   MSDDKPFLCTAPGCGQRFTNEDHLAVHKHKHE-----------------------------------------------------
ZDp:  ----KPFLCTAPGCGQRFTNEDHLAVHKHKHEMTLKF-----------------------------------------------
Zp:   ----KPFLCTAPGCGQRFTNEDHLAVHKHKHE-----------------------------------------------------
Dp:   ----------------------LAVHKHKHEMTLKF-----------------------------------------------
FXFP: ----------------------------------------------------EVGLFNELASPFENE-----
           Minimal JNK-binding region
```

Figure 47: Fine mapping of ATF2 region responsible for JNK1 binding with the help of GST pull-down. The segment encompassing both the Zn-finger and the predicted D-motif is both required and sufficient for JNK1 recruitment to the transactivator region of ATF2.

By comparing the two separate structures (ATF2 alone, where the N-terminus of the motif is rigidified, ATF2 peptide with JNK3, where only the C-terminus of the peptide appears to be in relevant conformation) with that of the pepNFAT4-JNK1 complex structure, I was now able to build a hybrid model, consistent with the presence of a Zn-finger. The similarity of amino acid arrangements between the motif of ATF2 and NFAT4 suggests a similar binding mode. The only major difference is that the 3-10 helix of NFAT4 would correspond to an alpha-helix in ATF2. A tighter helix, on the other hand

would be complemented by longer side chains in ATF2: Lys is found instead of an Arg and Met in the place of a Leu amino acid. Apart from holding the helix, the Zn-finger domain contributes suprisingly little to the binding surface in this model. But this could still influence affinities, as the disorder-to-order transition of alpha-helical linear motifs imposes an entropic penalty on binding energy. Previously, elegant examples demonstrated that artificial rigidization of alpha-helical motifs by synthetic crosslinkers ("staples") greatly increases their natural binding affinity.[177] In our current example, it is the most plausible that the Zn^{2+} ion provides a "natural stapling" structure, rigidifying an alpha helix in order to enhance its binding. This is also interesting from an evolutionary perspective. ATF2 is very distantly related to the Jun transcription factors. Their N-termini are also similar, including the transactivation motifs and the MAPK-binding elements. Jun proteins, however lack Zn-fingers: Instead they have classical NFAT4-like D-motifs that have to form a helix without the aid of external "staples".

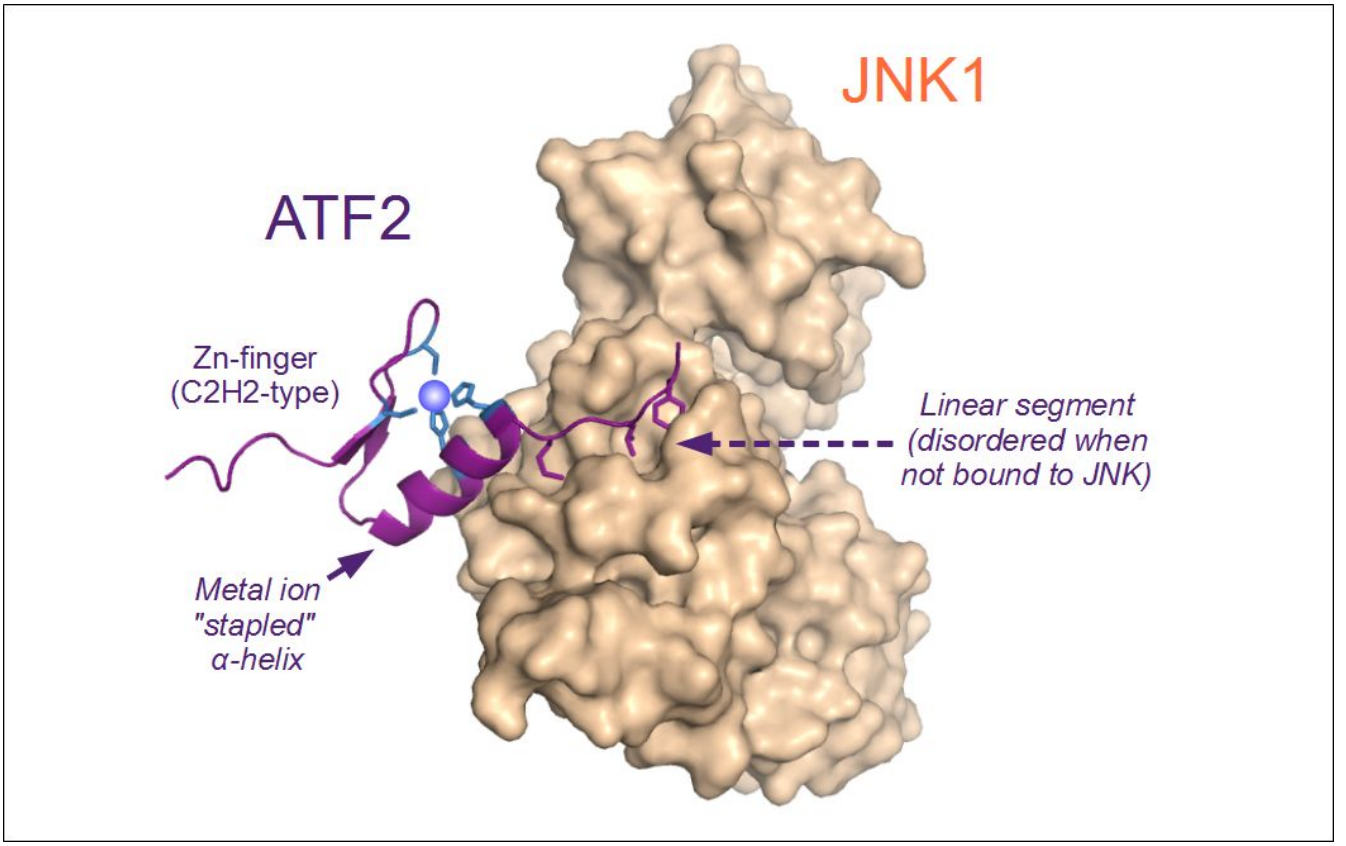


Figure 48: 3D model of how ATF2 is proposed to bind JNK1 (This was built from three separate, experimentally-determined molecular structures, the ATF2[19-56] solution NMR structure (PDB code: 1BHI) as well as the pepATF2-JNK3 (PDB code: 4H36) and pepNFAT4-JNK1 (PDB code: 2XS0) co-crystallized X-ray structures)

3) **The Rhodanese domain of MKP1**

Not all D-site binding elements are linear motifs. In the case of MAP kinase phosphatases (MKPs, also called Dual specificity phosphatases / DUSPs), the recruitment of the MAPK is mediated by a folded domain. Termed rhodanese domains due to their structural similarity to bacterial rhodanese enzymes, these domains are catalytically inactive. Their role is two-fold: First, they inhibit the dual specificity phosphatase domain in the absence of MAPKs. Second, they recruit MAPKs by docking in order to feed them into the catalytic domain for dephosphorylation. The MAPK-binding surface of rhodanese domains mostly maps to a single helix on the domain surface. The structure of a partial complex was determined by X-ray crystallography early on: this shows the helix in question from MKP3 binding to the D-site of p38α, similarly to genuine linear motifs.[109] But notably, the positioning of this single helix was not compatible with its partly buried position inside the intact MKP3 rhodanese domain (that was crystallized on its own).[178]

For a long time, it was unclear how the rhodanese domains can access the D-site of MAPKs. However, in 2011, the structure of the MKP5-p38α complex was successfully solved, with an intact rhodanese domain.[138] A rather different arrangement here indicates that the structure of the partial MKP3-p38α complex was possibly an artefact. In fact, rhodanese domain-MAPK interactions act like typical domain-domain interactions. A discontiguous surface is responsible for the recruitment of MAPKs, with sequence-wise distant amino acid side chains coming together to mimic a linear motif. The arrangement of side chains are indeed remarkably similar to a long reverse D-motif like that of MAPKAPK2, known to bind p38α and ERK2 (with a lower affinity), but not JNK1.

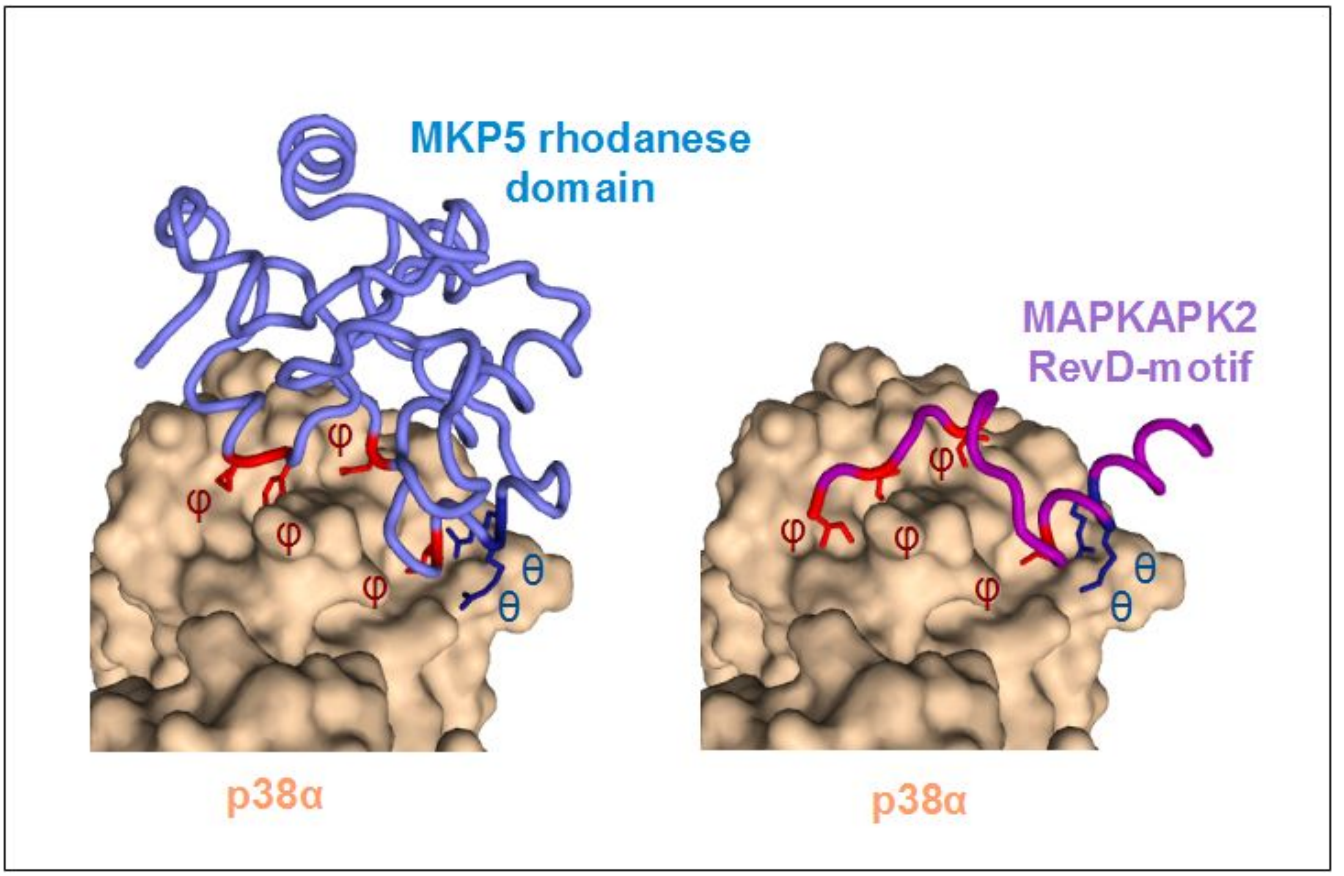


Figure 49: The folded rhodanese domain of MKP5 (blue) binds to p38α (beige) in highly similar manner to a genuine reverse D-motif of MAPKAPK2 (magenta). Even the contact points (φ and θ) correspond to each other, despite being arranged in a non-contigous manner on the rhodanese domain. (This figure was created from the X-ray structures with pdb codes: 3TG1 and 2OKR)

But MKP5 is just one member of the large and diverse MKP family. Another member, MKP1 has a similar domain architecture, but shares relatively low sequence homology to MKP5. To study the binding profile of MKP1, I successfully cloned and expressed the rhodanese domain of MKP1. The MBP pull-down shows that MKP1 binds with high affinity to ERK2 and p38α, but at a somewhat lower extent to JNK1. Subsequent fluorescence polarization (FP) measurements conducted by my colleague, Ágnes Szonja Garai more clearly demonstrated that the rhodanese domain of MKP1 associates with the docking site of ERK2 and p38α only (Kd=0.2μM for ERK2 [against CF-pepHePTP], Kd=3.3μM for p38α [against TAMRA-pepMEF2A] and no binding detected with JNK1 [TAMRA-pepNFAT4]). This is well in-line with the already-described catalytic properties of the full MKP1 enzyme: capable to dephosphorylate all three MAPKs, but requiring the rhodanese domain only in case of ERK2 and p38α.[179] Due to the architecture of rhodanese domains (i.e. that they all contain a CD-groove contacting helix), it appears that they are structurally unsuitable for strong interaction towards JNK (that has no

CD-groove). However, this contradicts the observation that many MKPs, most prominently MKP5 are capable to dephosphorylate JNK1 (in addition to p38α) as their preferred substrate. During our motif finding studies, we did identify a JNK-binding motif in MKP5, separate from the rhodanese domain, which could be the natural solution to this problem.

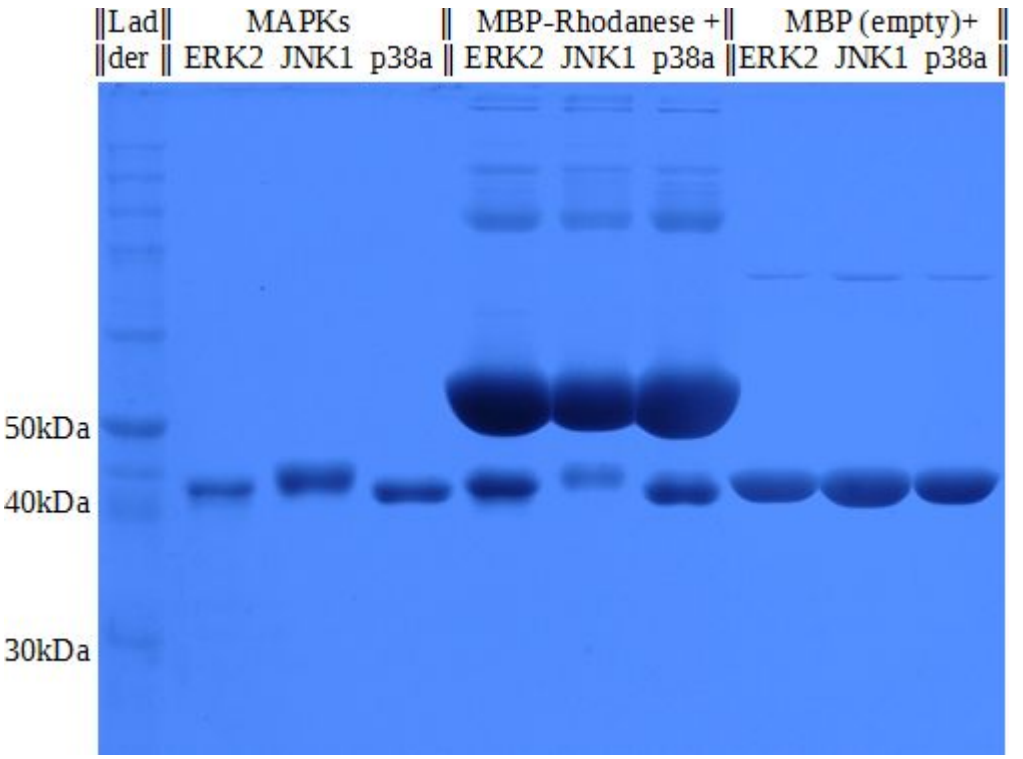


Figure 50: MBP pull-down experiment with the (MAPK-recruiting) rhodanese domains of the phosphatase MKP1. The rhodanese domain appears to bind ERK2 and p38α much stronger than JNK1.

DISCUSSION

Limitations of MAPK partner identification through systematic modelling

In the current study, it was demonstrated that canonical, D-motif dependent partners of MAPKs are in fact quite common - if one knows where to search for them. This was made possible by a set of improved consensus sequences. These, in turn, were the results of a complex procedure that combined sequence and structure based bioinformatics and experimental validation. Though powerful, the method still does not find all MAPK binding proteins, therefore it can only be used to identify a "representative subset" of MAPK interactomes. A number of partners with atypical or **"naturally defective"** docking motifs do exist (e.g. MKK5, MKK3, TAB1), which are difficult to predict. Such defective motifs often act in a non-autonomous way: these weak elements are frequently complemented by additional protein stretches, motifs or domains. Besides, not all MAPK binding elements are linear motifs. **Folded domains** such as the rhodanese domain of dual-specificity phosphatases binds to the same site as intrinsically disordered docking motifs. It should be noted that motifs other than the canonical D-motifs (e.g. the so-called FxFP motifs) also exist. A considerable number of interactions might also be indirect, mediated by a third partner. Nevertheless, directly interacting with a MAPK **solely through short linear motifs** appears to be a major and extremely widespread phenomenon in mammals.

Some of the newly-identified partners directly fit into the **core of MAPK pathways**. These include specific phosphatases as well as MAPK kinase kinases (MAP3Ks). While there can be little doubt that docking motifs of phosphatases would be required for MAPK dephosphorylation, the presence of docking motifs in MAP3Ks is a more intriguing observation. It is probable that phosphorylation of proteins acting on the MAP3K level (like on MEKK1, MLK1/2 or KSR2) would allow direct feed-back control of MAPK pathways - as it was described for the mammalian KSR1 or the yeast protein Ste5. However, majority of novel hits appear to lie outside the core MAPK pathway module, and they are probably simple downstream elements (i.e. substrates). Unfortunately, the methods applied in my studies are not sufficient to verify enzyme-substrate connections in cells. Most of the novel proteins are still expected to be either **direct MAPK substrates** or scaffold proteins (i.e. enabling phosphorylation of **indirect MAPK substrates** through protein complexes).

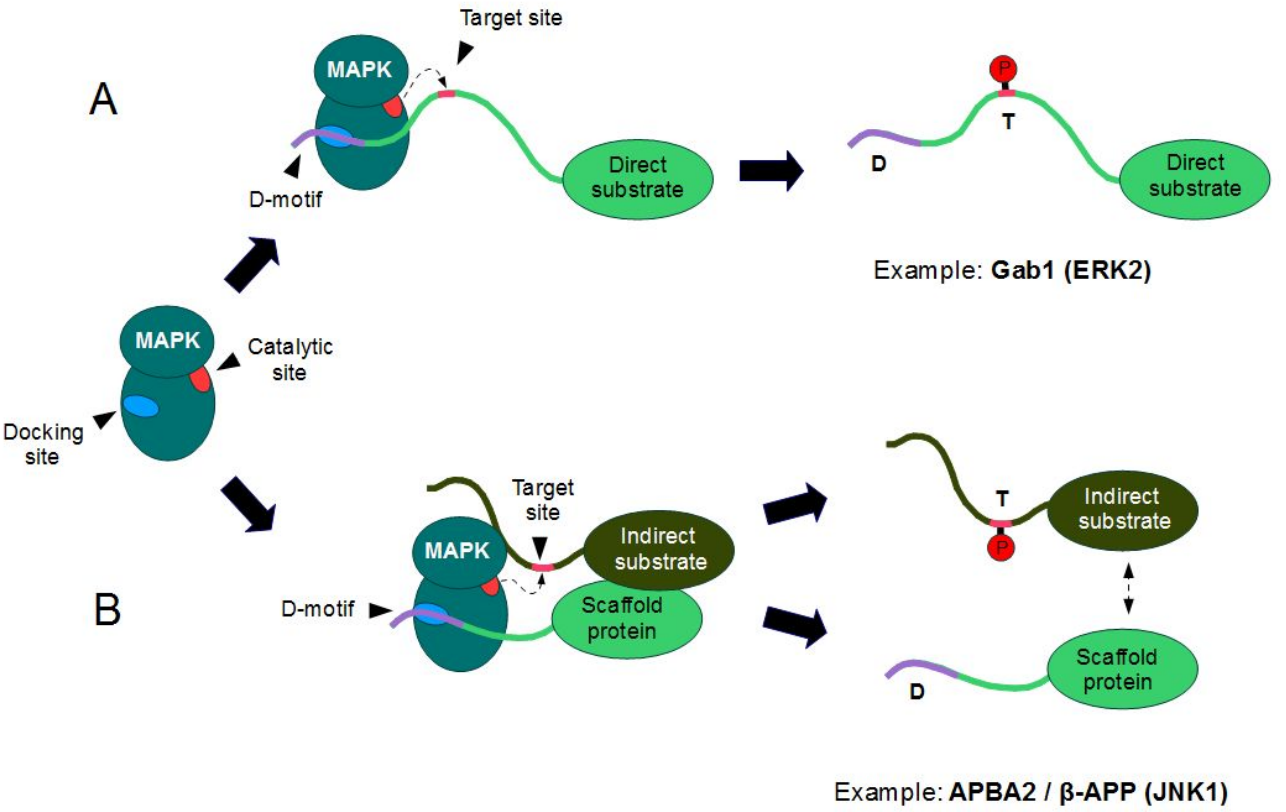


Figure 51: Schematics of a MAPK acting on its direct substrates (A) or indirect substrates (B), using D-motifs of either the substrate itself, or a middle protein associating with the substrate.The systems cited as examples (ERK2/Gab1 and JNK1/APBA2/beta-amyloid precursor protein) are already well-known as substrates, but their D-motifs were first identified in the current study.[171,180]

Of all the MAPKs studied, ERK2 is the most widely explored by previous research. Several different methods were utilized to identify ERK2 substrates by **large-scale phosphoproteomics**. [181–183] Unfortunately, pairwise overlaps between the lists of substrates are usually low across studies (around ~10%, see Courcelles et al, 2013), illustrating their unreliability and the sharp dependence of substrate identification on the experimental conditions used.[183] A comparably low degree of overlap can be detected between our predicted substrates and those detected by large-scale phosphoproteomic studies. It was, however noted that D-motif like sequences are enriched in experimentally detected ERK2 substrates, yet detection or verification of direct physical association was not performed.[182] Those studies that used more-or less high throughput methods (e.g. yeast two-hybrids) to identify partners of JNK1 or p38α by direct physical interaction, usually also resulted in a very low number of hits (possibly reflecting limitations inherent to the methods).[174,184] Thus it is very clear that

the direct physical enzyme-substrate interactions of even the best known MAPKs were never mapped to a satisfactory degree.

A new paradigm of MAPK-dependent regulation of substrates

For many decades, up until the early 2000s, most of the phosphorylation events in protens were believed to induce a **conformational change** in the tertiary structure. The same was thought of the MAPK-dependent phosphorylation of critical transcriptional regulators, including Elk-1, c-Jun or NFATs. However, these theories were born many years before intrincically disordered proteins were discovered. It was demonstrated by CD spectroscopy in an elegant experiment, that concomitant phosphorylation of many (potentially >10) sites on the N-terminal segment of NFAT2 cause no detectable structural change.[185] This has been a great puzzle, as previous models envisioned a conformational change by which the nuclear localization signal (NLS) motif of NFATs could be masked upon phosphorylation. However, this was a false assumption: despite earlier speculations, the phosphorylated motifs do not appear to be able to interact with the NLS motif directly.

In the recent years, it has become increasingly clear that serine/threonine phosphorylation events mostly take place on **intrinsically disordered** segments.[186] The only case when domains themselves are phosphorylated, happens when the site in question is located on a fairly mobile “loop”. The situation is thus somewhat different from Tyr phosphorylation, where the rather long side chain of tyrosine amino acids enable them to be phosphorylated even on the surface of relatively rigid domains.[187] But even in the case of tyrosine, the majority of in vivo detected phosphorylation events are still located to disordered segments. The generic connection between phosphorylation sites and intrinsic (local) disorder also applies to MAPK target sites. Therefore the traditional model of phosphorylation-induced conformation changes is practically untenable. We must assume that in the overwhelming majority of cases, MAPK target sites themselves form linear motifs, with the phosphorylation either enhancing or disrupting some inter- or intra-molecular interaction. Such mechanisms are called “phospho-switches” in the literature, but unfortunately very few examples were described for MAPKs.

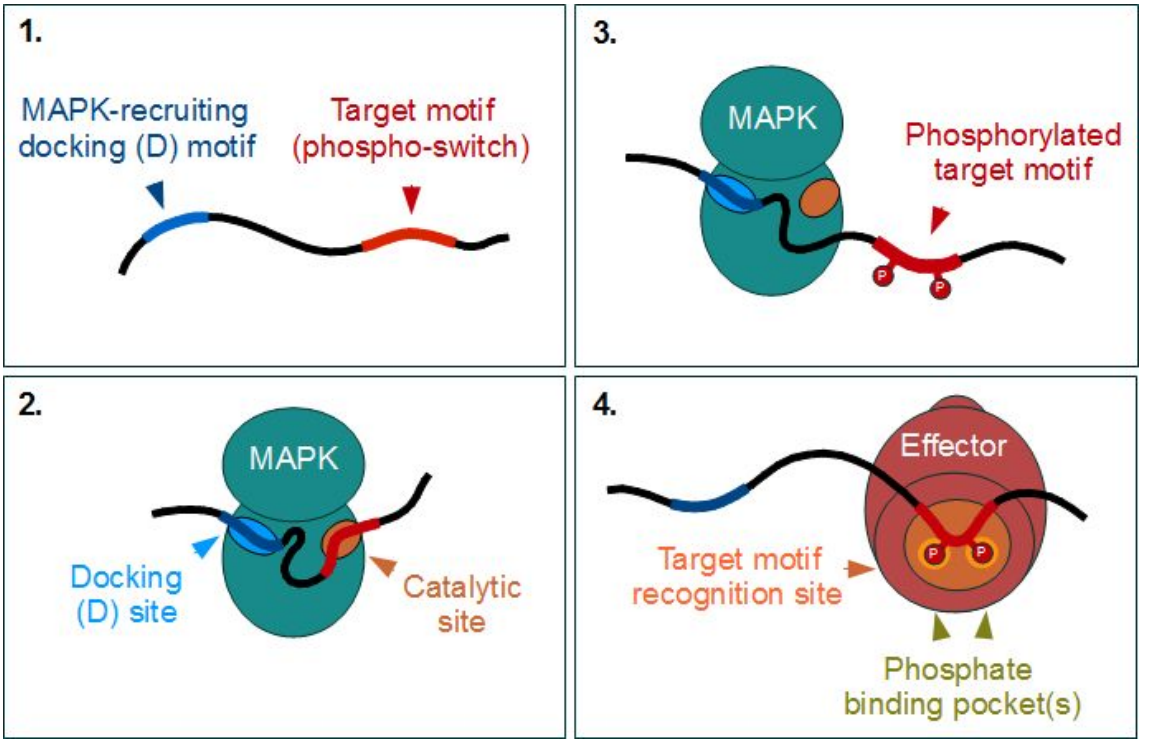


Figure 52: A proposed generic mechanism for MAPK-dependent phospho-switches on target proteins (numbers 1 to 4 indicate the sequence of events)

My assumption that MAPKs would mostly exert their effects by regulating protein-protein interactions through specific linear motifs, gains reinforcement from a series of simple observations, on known MAPK substrates (see figure). The first of these examples relates to two separate families of bZIP transcription factors. The Jun and ATF2 families are both regulated by JNK-dependent phosphorylation events, although they recruit their partners through structurally rather different elements. In the case of Jun proteins, JNK binds to an NFAT4-type docking motif located near the N-terminus of c-Jun (also found in its relatives, JunB or JunD). In the case of ATF2, recruitment of JNK mainly depends on a Zn-finger instead, that mimics an NFAT4-type docking motifs. However, if we look at the phosphorylation target motifs (tandem phosphorylation target sites with their surroundings) controlled by JNK in the two divergent protein families, they turn out to be strikingly similar. The importance of phosphorylation on these sites for the control of transactivation was known for decades: Unfortunately, we still know very little of its interacting partners.

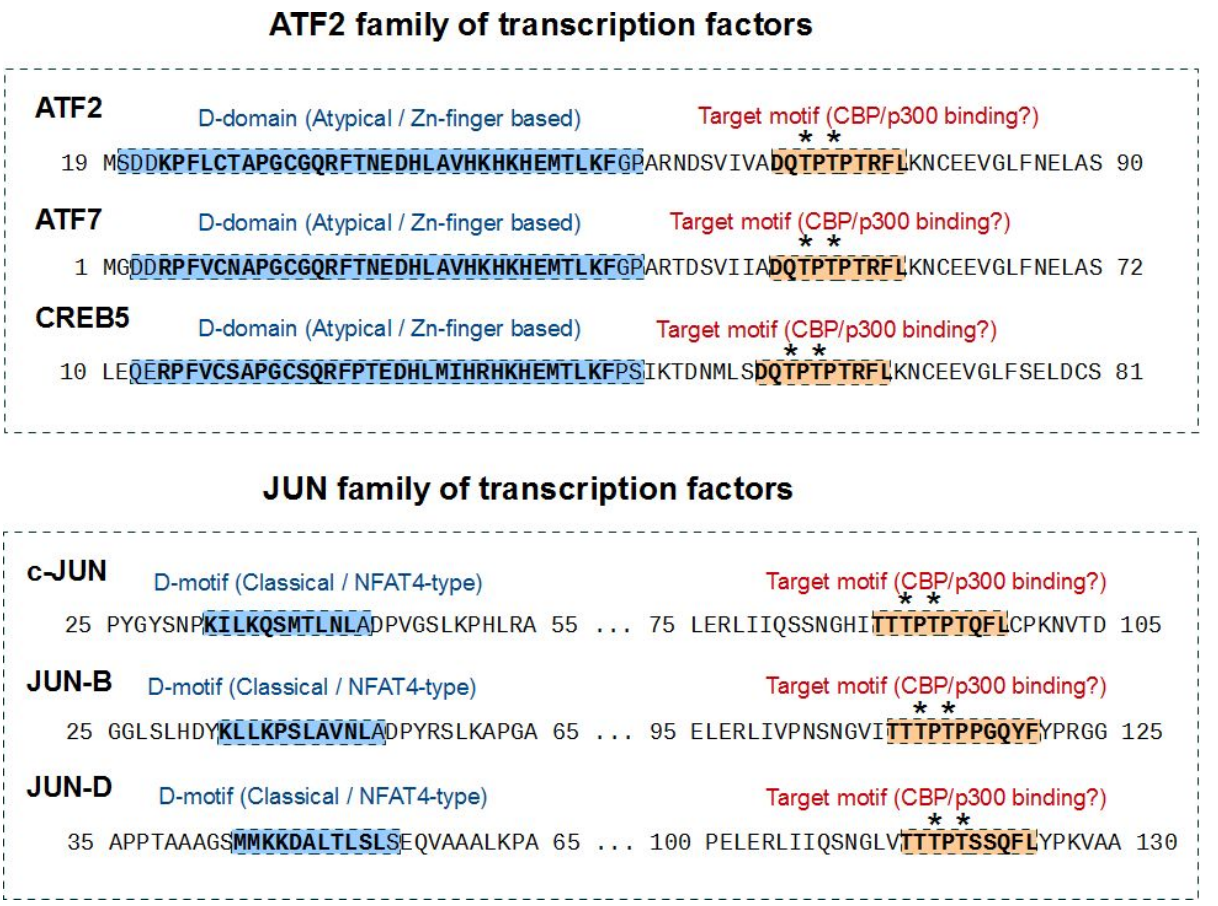


Figure 53: Comparison of docking elements and phosphorylation target motifs in the ATF2 and JUN family of transcription factors

A very similar phenomenon can be seen when comparing the surrounding of critical phosphorylation sites in MEF2A (also found in MEF2C and MEF2D, but not in MEF2B) with that of c-Fos (the same site is found in Fosl2 but not in FosB). Despite the fact that the two protein families are utterly unrelated, bind the DNA through structurally different domains, and even recruit the associated MAPKs through different motifs and sites (MEF2A has a classical D-motif, while c-Fos has an FxFP motif), the target sites are surprisingly similar. Not only the arrangement of tandem SP/TP sites, but the intervening amino acids also correspond to each other. Again, the motifs are too special for this pair to be a random coincidence, and suggest convergent evolution instead. Both sites are known to confer phosphorylation-dependent transactivation to the targeted protein, yet the exact proteins binding these sites are still completely unknown. The motifs do show some distant structural similarity to a histone H3-binding motif found in the protein HJURP.[188] What we know is that the co-evolution of target sites with D-motifs strongly suggest a “phosphoswitching” function. In both the MEF2 and the Fos families, some vertebrate paralogs tended to loose both the target sites and the motifs. We can be almost certain

that this is a secondary loss, since non-vertebrate genomes that carry only a single copy of either gene typically tend to conserve the whole regulatory module. Notably, both the target motifs and the recruitment motifs display an island-like conservation pattern.

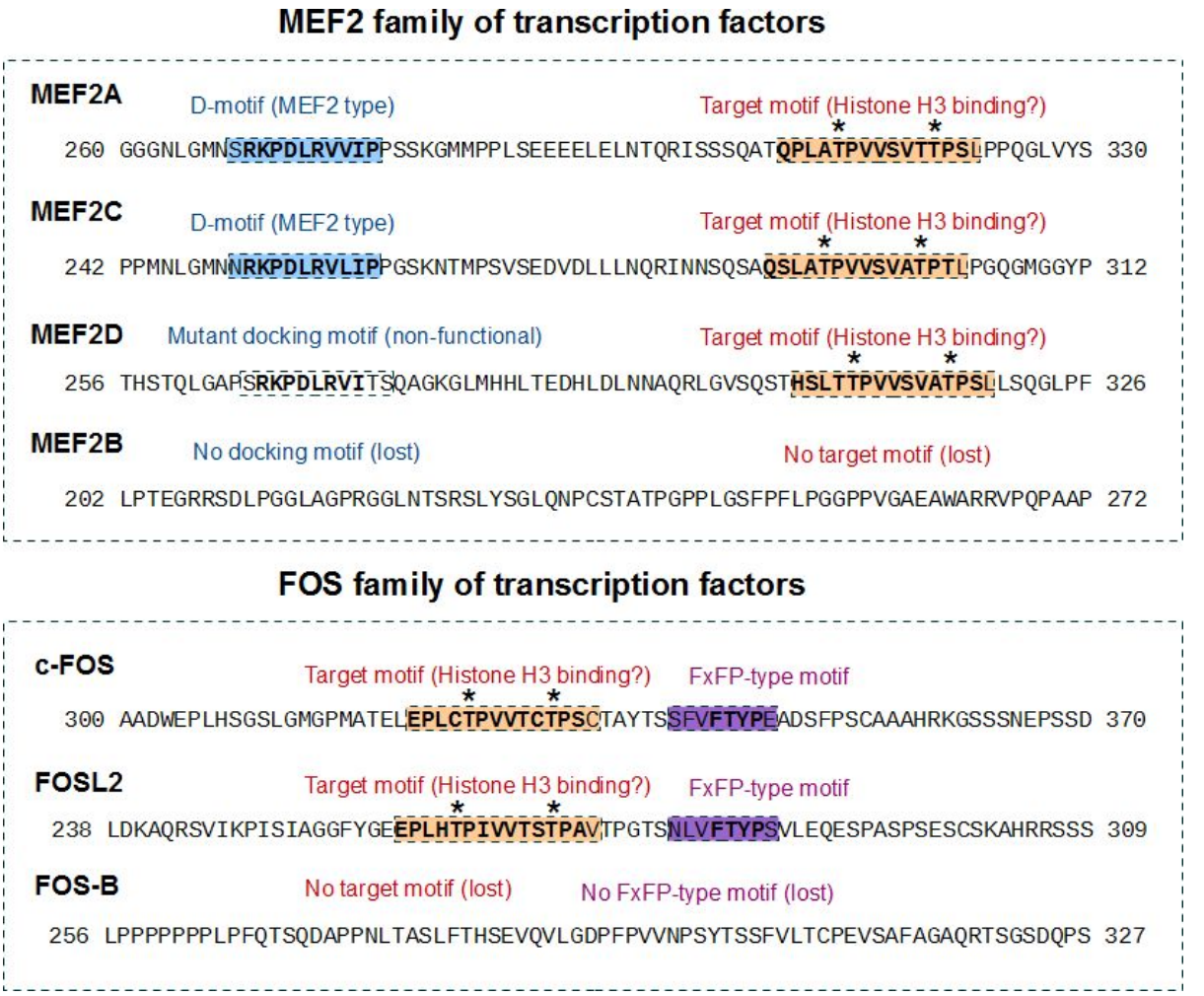


Figure 54: Comparison of the docking motifs and target sites in the MEF2 and FOS family of transcription factors. Note the high amino acid similarities around the phosphorylation sites.

If some linear motifs controlled by MAPK phosphorylation are non-unique, they should be (at least theoretically) identifiable from a set of non-related substrate proteins, as similar or identical, phosphorylatable motifs. Although development of such a method was well beyond the scope of the current study, we did apply a preliminary analysis to our newly-identified partner proteins . A comparison of them with each other and other known partners suggest that a particular motif ([ST]PPx[ST]P) occurs rather commonly in the neighbourhood of D-motifs. Fortunately, some instances of the cited motif were arleady positively identified as a recruitment element for the E3

ubiquitin ligase FBW7. Related elements could be found in a large number of predicted MAPK partners (substrate candidates), lending credence to the hypothesis that FBW7-dependent (K48-linked, degradative) ubiquitinylation of substrates might be frequently elicited by MAPKs. This would not be an entriely novel concept, as MAPK-dependent phosphodegrons were already identified in yeast Tec1 transcription factor (utilizing the E3 ubiquitin ligase Cdc4, which is the Saccharomyces ortholog of FBW7) as well as in human c-Jun and a number of other proteins. [160,189–191]

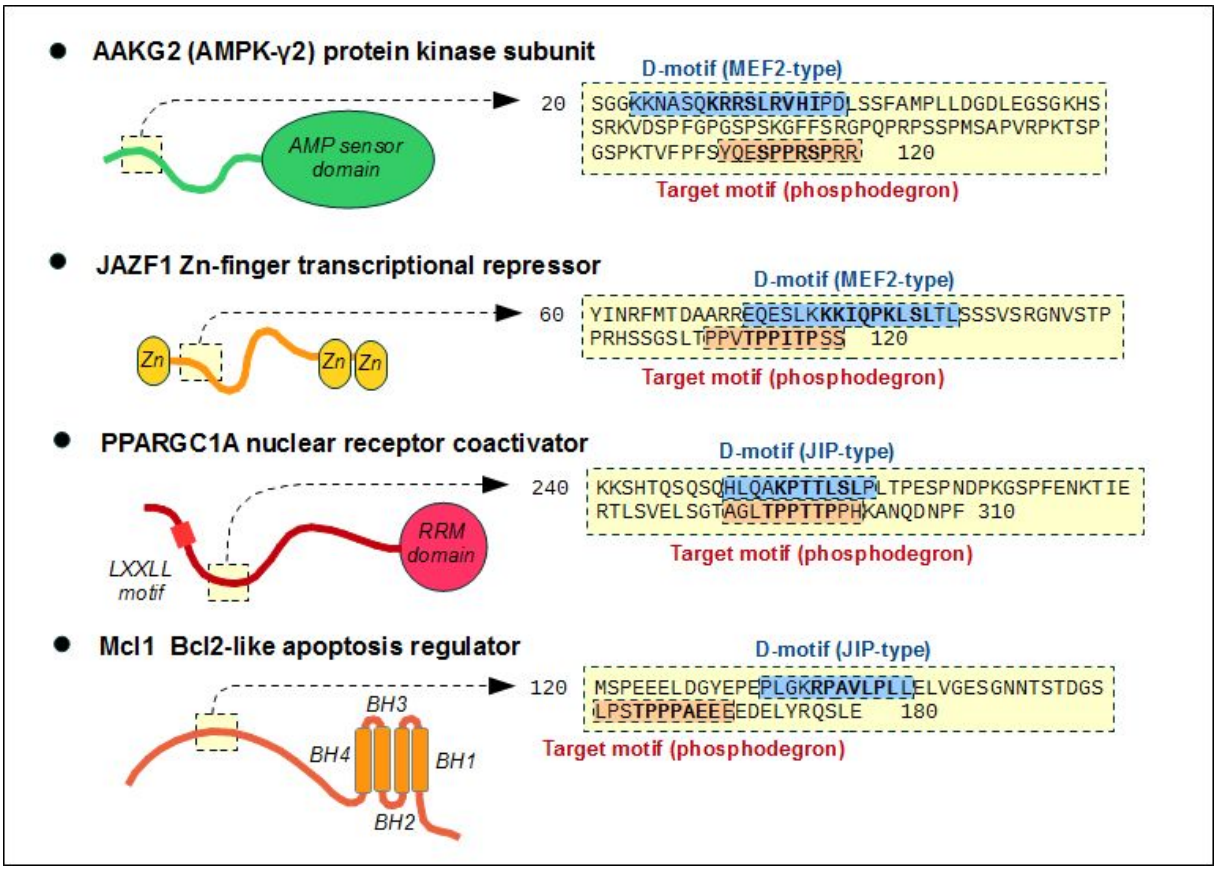


Figure 55: Newly-identified MAPK partner proteins (detected in the current study), displaying similar, phosphorylatable motifs downstream of their docking motifs - that are all potential FBW7 phosphodegrons (In the case of PPARGC1A / PGC1A and Mcl1, the degron has already been validated experimentally by independent research groups)[190,191]

Implications for systems biology and evolutionary biology

Our experiments and subsequent predictions gave an exquisitely detailed, **molecular-level map** of MAPK pathway targets. They support the notion that specific human MAPKs regulate diverse physiological processes. Many of these roles have already been identified by organ-level or whole-animal based experiments (e.g. knockout models). However, these regulatory roles were previously often attributed to single target proteins. For example, the role of JNK in axonal growth was attributed to the JNK-JIP1 interaction, and the association of JNK with diabetes was attempted to be explained by the JNK1-IRS1 interaction alone. In contrast, our results strongly imply that these interactions are just the tip of the iceberg: JNK (as all MAPKs) connects to the targeted physiological systems by a large number of direct interactions. Whereas individual connections might not be stable (especially in an evolutionary sense), having so many specific linkages could provide a robust physiological regulation.

Surprisingly, many of the newly implied MAPK partners have a **restricted expression pattern**, enabling fine-tuned regulation in specialized tissues. Because of the latter phenomenon, a great deal of these interactions are unlikely to be discovered by large-scale protein-protein interaction screens. Easy-to handle cell lines and mass-spectrometry based analyses provide a powerful tool, but not for proteins that are only expressed in special, differentiated tissues (e.g. AAKG2, which is only abundant in cardiomyocytes) or restricted to certain embryonic developmental stages (e.g. DCX is almost exclusively expressed in developing neuroblasts). Here, a modeling-driven interactome search is the most suitable tool to fill in the gaps in our knowledge.

A comparative analysis suggests that rapid changes happened to MAPK pathways during the **early evolution of vertebrates**. Target proteins could have been brought under tight MAPK control by simply introducing docking motifs, however, this also necessitated target sites where phosphorylation could elicit functional effects. Similarly, existing links could have been thrown away by the loss of docking mechanisms. This process was apparently the fastest when early vertebrates diverged from other chordates (that might or might not have coincided with a paleobiological event known as the “Cambrian explosion of life”, when evolution of multicellular animals was extremely fast). Since then, the rate of motif emergence seemingly slowed down, but it has not completely stopped (some novel human motifs could be traced back only to mammals).

My results regarding the fast evolution of MAPK docking motifs can probably be explained by the ease of *de novo* assembly of **MAPK regulatory modules**. It is highly likely that by a mere coupling of two short linear motifs (a docking motif and a target motif) on a disordered segment, MAPK-dependent regulation can be simply introduced onto any protein. The fact that MAPKs have loose primary substrate specificity is very significant in this regard: It enables MAPKs to couple with a rich variety of target motifs, each controlling the binding of different proteins, depending on their phosphorylation state. This gives way to many intriguing evolutionary scenarios: Obtaining *de novo* MAPK-dependent regulation by simultaneous evolution of docking and target motifs; swapping the kinase partner in an existing module by creating a new docking motif to a pre-exiting target site or changing the effect that MAPK phosphorylation has on the protein, by introducing a new target site. Since our insight on MAPK target sites is much more limited than what we have uncovered of the docking motifs, we are definitely en route to many intriguing discoveries in the future years!

MATERIALS & METHODS

In silico search for motif candidates and filtering of improper hits

The Human Proteome Database (HPD) contains 20,248 sequences. The HPD was scanned for motif hits with basic pattern matching using the regular expressions, yielding 81,722 hits accross the 7 motif classes/types (JIP1-class, NFAT4-class, MEF2A-class-MEF2A-type, MEF2A-class-MKK6-type, DCC-class-generic, HePTP-class-Ste7-type, HePTP-class-HePTP-type). The estimation of the interaction potential of the selected protein regions was done with the ANCHOR algorithm, a method trained to recognize binding regions in disordered protein segments. In linear motif selection a more permissive version of ANCHOR can be used, therefore the default 0.5 cutoff value was lowered in an adaptive way. Motif hits were kept only if they overlapped with either an ANCHOR region predicted by using a 0.4 cutoff, or an ANCHOR region predicted by using a 0.3 cutoff, but in this case at least one of the 20 residue flanking regions of the motif hit had to have a sufficiently high average disorder value (above 0.45) predicted with IUPred. As a result the number of hits was reduced to 19,113.

Motif hits were discarded if they resided in proteins that were predicted to have a signal peptide by signalP and if they were also predicted not to have a transmembrane region predicted by Phobius. These motif hits were predicted to be localized outside of the cell, which is incompatible with MAPK binding. Phobius alone was also allowed to predict signal peptides alone, if signalP score were not too low (above 0.3). If a motif instance resided in a protein that was predicted to have a signal peptide but it also had at least one transmembrane region, the localization of the motif region was further checked. If it was entirely intracellular, it was kept, otherwise discarded. This filtering step reduced the number of motif hits to 17,082. All hits that were predicted by WolfPsort to be extracellular (with score >=25), membrane protein (with score >=25), localized to the E.R. (with score >= 15) or the Golgi (with score >= 9) were filtered out, unless they harbored transmembrane regions and the region containing the motif was predicted to be localized in the intracellular space. There were 16,805 hits remaining after this step. Motif hits that were determined to reside in Pfam domain regions were discarded. Hits were also omitted if they coincided with Pfam Family/Repeat/Motif regions but were in a manual curation process deemed to have a stable structure in isolation which is incompatible with localization in disordered protein regions. Furthermore, motif occurrences that overlapped with coiled-coil regions predicted by COILS were removed as well. Finally there were 12,435 motifs remaining for further analysis including more than 90 % of the known positives.

Template building, complex modelling & scoring by FoldX

Each major structural class had at least one experimentally determined model available in theProtein Data Bank. These were modified and also re-fitted when necessary. Sections of the peptides that fell outside the consensus motif were removed in all X-ray structures, with only a pair of amino acids immediately flanking the motif left in place (in addition, these were always mutated to alanines). These modifications were done in the software COOT, with the help of Gergő Gógl. The JIP1 and the NFAT4 classes were modelled from the JIP1-JNK1 and NFAT4-JNK1 complexes, respectively. For the greater MEF2A class, we used the MEF2A-p38α and MKK6-p38α complexes. However, the latter two X-ray structures were were both incomplete: Only a few amino acids were visible from the linker towards the N-terminus, and the charged N-termini were mostly or completely missing (had poor electron density or none). Therefore we removed the complete N-termini of these motifs (these were deemed unreliable), and used only the hydrophobic part for modelling. The ERK2-pepHePTP complex contained and artificial disulfide bond at the φ_B position that facilitated crystallization of this protein-peptide complex, however it also distorted the conformation of the docking peptide at the C-terminal region. We had to delete the artificial disulphide bridge in the complex and susbtitute the last Cys of the peptide with a Val at the φ_B position, in order to obtain a model of the native complex. In addition, we submitted the complex to a Rosetta-based optimalization (using the PepFlexDock server, with no restraints), to correct distortion due to the forced covalent cross-linking. The model for the hypothetical Ste7 model was constructed from the yeast Fus3-pepSte7 and Fus3-pepMsg5 structures. In complex with the yeast MAPK Fus3, the Ste7 peptide displays a somewhat shorter N-terminus than the motif of Msg5. However, this merely appears to be a crystallization artefact: the former complex was determined with a shorter peptide than the physiologically relevant motif (confirmed by the presence of conserved hydrophobic amino acids at the position corresponding to that of Msg5). To obtain a model for the ERK2-pepSte7 complex, we first rebuilt the full Ste7-Fus3 complex based on the Msg5-Fus3 one, then superimposed it on human ERK2. As the last step,we optimized the resulting theoretical complex with PepFlexDock. For the DCC class, we had no separate model for the Far1 subtype, as all available structures on human proteins are representing the DCC subtype. However, these two are probably similar enough (only differing in the placement of prolines not contacting the MAPK surface directly) to be modelled by the same template. The only problem with the DCC-ERK2 complex was that we had to remove the N-terminus of the peptide similarly to many previous ones, as most of it was missing or poorly supported by electron densities in the X-ray structure.

Complexes of peptides in question with a MAPK were modelled using FoldX, similarly as it was previously described for SH2-binding peptides. This step was performed by Tomas Bastys and Olga Kalinina at the Max-Planck Institut für Informatik, Saarbrücken. For each selected motif, the corresponding model was mutated in its sequence, and then the resulting raw structure was energy-minimalized using the RepairPDB function of FoldX. The scoring was run in batches on a local cluster. The estimated interaction energy of the complex was calculated for each motif, from the sum of several elementary terms (backbone H-bonds, sidechain H-bonds, Van der Vaals interactions, electrostatics, solvatation of polar segments, solvatation of hydrophobic segments and Van der Waals clashes). For the final interaction energy, an additional “peptide stability in solution” correction term was also taken into account. The obtained interaction energy+stability values were used to rank motifs according to their “goodness” for initial testing.

Production of inactive and phosphorylated MAPKs

MG950 bicistronic plasmids encoding for dephosphorylated MAPKs with lambda phosphatase were the same as the ones used in earlier studies.[31] Plamids for the production of activated MAPKs were built similarly, but with the appropriate MAP2K in place of the lambda phosphatase (MEK1 for activated ERK2, MKK7 for activated JNK1 and MKK6 for activated p38α) by Ágnes Szonja Garai, Tünde Bárkai and Gergő Gógl. These constructs encode full-length MAPKs (except for JNK1, where the shortest natural isoform JNK1α2, was used as "inactive JNK1" and its C-terminally truncated version ΔC20JNK1α2 for generation of "active JNK1"), with an N-terminal His_6 tag. All MAPKs were expressed and purified using standard methods (see the appropriate chapter below). While the activity of p38α was sufficient after co-expression with MKK6 and subsequent Ni-NTA and ion exchange purification on a resource-Q column (GE healthcare), we had to resort to additional steps in other cases. Activated ERK2 was additionally purified on a mono-Q column (GE healthcare) by Gergő Gógl to separate differentially phosphorylated forms (and monitored by SDS-PAGE followed by ProQ diamond staining). Generation of sufficiently activated JNK1 required an extra in vitro phosphorylation by the appropriate, separately purifed MAP2Ks, and subsequent separation by ion exchange (this was done by Imre Törő).

Design and construction of synthetic substrates

The synthetic MAPK substrates used in the phosphorylation assays were encoded by a single vector family (pAZAD). This was built from a modified Novagen pET vector (pETARA, derived from pET-17), encoding N-terminal GST and C-terminal His6 tags. pAZAD also incorporates a series of flexible linkers, a substrate site derived from human ATF2 (lacking the T69 site, with only the T71 site present) as well as a D-motif containing cassette. As the first step, a pair of synthetic oligonucleotides encoding the sequence GSGADQAPTPTRFL (linker+substrate) were insterted through the cloning sites NotI and XhoI. Subsequently, the vector was digested with BamHI and NheI, and a new, upstream linker was inserted, separating the future D-motif containing site from the upstream GST tag. The upstream linker encoded the sequence ADQAPAPARFLGSGRGS. This is similar to the downstream linker+substrate sequence, but lacking MAPK-phosphorylatable amino acids (we made use of the fact that the corresponding segment of ATF2 was shown to be intrinsically disordered by NMR spectroscopy). The upstream linker was inserted using a BglII (compatible with BamHI) and NheI sites, creating an upstream BamHI/BglII hybrid site and a downstream NheI site. However, the inserted sequence also contained a perfect BamHI site immediately upstream of the NheI site, thus moving and reconstituting the BamHI site of the original vector at a different location. In the end, the original insert between the BamHI and NotI sites, mostly inherited from the multicloning site of pETARA (encoding the protein sequence ASAVDLVPRG) was cut and exchanged for the appropriate D-motif containing fragments. The linker from the D-motif to the phosphorylatable amino acid was designed to be at least 12 amino acids long. Previous observations indicated that >9 amino acids are needed to be inserted between the D-motif and the target site for the coupling to be effective; This was also confirmed by the inspection of MAPK-D-motif complex structures as well as a meta-analysis of known substrates.

Ligation of synthetic oligonucleotides into the target vector

The oligonucleotites destined for direct ligation were typically designed as a “virtually digested” complementary pair, with sticky ends corresponding to the endonuclease cleavage sites in case of constructs shorter than 60 nucleotides. Whenever a construct of >60 nucleotides was required, it was built from four overlapping oligos. The lengths of these segments were designed carefully so that each of them would overlap with the complementary strands on similarly long stretches (avoiding radically different annealing temperatures). Their sequence was generated by reverse-translation using the SMS2 online tool (http://www.bioinformatics.org/sms2/rev_trans.html) and primarily codon-optimized for expression in *E. coli* (K12 strain), by the Codon Usage Database (http://www.kazusa.or.jp/codon/). As with PCR primers, these oligonucleotides were also checked with the IDT Oligo Analyzer (http://eu.idtdna.com/calc/analyzer) for the presence of loops with high Tm (within 10ºC of annealing Tm) as well as for extensive self-complementarity (>10kcal/mole). Whenever potential problems were detected, individual nucleotides were modified (while keeping their pre-defined amino acid sequence) until the predicted, deleterious nucleotide features were corrected. All synthetic oligonucleotides were ordered from Integrated DNA Technologies (IDT), on the smallest scale of synthesis available.

The delivered oligonucleoties (typically around ~30nmole in each tube) were dissolved in doubly distilled (sterile filtered & autoclaved) water (ddH2O), for an end concentration of 100µM. For the prupose of annealing, 5µl of each oligonucleotide were added to a mixture of 10xT4 ligase buffer (30µl) and doubly-distilled water (260µl in case of a pair, 250µl in case of a tetrade of oligos) for an end volume of 300ul. The annealing mix was heated rapidly to 95ºC in a programmable thermostat. Thereafter it was left to cool slowly over a period of 1-2 hours until room temperature was reached. At the next step, portions of annealed oligonucleotides were phosphorylated at their 5' ends by an in vitro reaction. Typical conditions were the following: 2ul annealed sample was added into a reaction mixture containing 15µl ddH2O, 2µl T4 ligase buffer (10x) and 1ul polynucleotide kinase (PNK, Thermo Scientific). The reaction was performed at 37ºC for 1 hour, after which the PNK enzyme was inactivated by incubation at 65ºC for 20 minutes. The synthetic oligonucleotides were still too concentrated to allow ligation into the vector at an 1:1 ratio, therefore samples of PNK-treated oligonucleotide mixtures were diluted by a factor of 10 before the ligation. The latter reactions were usually performed in a volume of 10ul, containing 1ul doubly-digested, phosphatase-treated & purified plasmid solution, 1 ul oliginucleotide mix (/10 dilute), 1µl T4 ligase buffer and 0.5µl T4 DNA ligase (Thermo Scientific) as well as 6.5µl ddH2O. Ligation reactions were done at room temperature, typically allowed to proceed for 2-3 hours before the constructs were transformed into KCM competent

XL1-Blu *E. coli* cells, using standard procedures for transformation. Colonies grown on plates containing appropriate antibiotica were counted against those from the control reaction (digested vector only, no insert). Colonies were selected for subsequent plasmid preparation only when the number of colonies was higher than the control (typically >5 times higher, unless the insert was constructed from 4 overlapping oligonucleotides, where ratios were often lower, therefore a ratio of >2 was deemed acceptable).

The selected colonies were marked and used to inoculate 3ml volumes of TB media (supplemented with appropriate antibiotics); these were grown overnight and subsequently used for plasmid preparation (using the PureLink DNA miniprep kits from Invitrogen, without gross modification to its protocol). Samples from the plasmids, eluted in 50µl buffer were sent for validation via Sanger sequencing (Biomi Kft, Gödöllő) in each case. As we have found that constructs built from synthetic oligonucleotides are relatively prone to contain errors (especially where they contained multiple GATC or GATT or similar sites, presumably due to interference with bacterial mismatch repair), all constructs that were used for protein expression had to be validated by at least a unidirectional sequencing.

Cloning of protein fragments from full-length clones and cDNA libraries

The full-length clones and long fragments used for *in vitro* experiments with DOCK5, CrkII, ATF2 and MKP1 were generated by me through PCR from a cDNA library prepared by Anita Alexa from HEK-293 cells. Longer Pbs2 constructs were generated by PCR from a full-length yeast Pbs2 clone. The wild-type constructs (AAKG2, RHDF1, KSR2, MKP5, DCX, APBA2, FAM122A, etc.) used for in-cell bimolecular fragment complementation (BiFC) assays were produced similarly from a cDNA pool by Ágnes Szonja Garai and Anita Alexa. Mutants of those constructs were either created by quickchange mutagenesis (in case of point mutations) or by shortening them with subsequent PCR reactions using a different primer at one end (in the case of truncations, used for AAKG2 and MKP5 ΔD-motif versions). For most reactions, we used the Phusion DNA polymerase in PCR reactions, and a common protocol with elongation times adjusted to the length of the construct to be generated (see table). In a few select cases, where PCR was unsuccessful using the base method, we used either the additive DMSO (up to 4%) or an extra pre-denaturation step (before the dNTP or the polymerase has been added).

PCR mix		
Components	ddH2O	13.5µl (ad 25µl)
	5x GC buffer	5µl
	Fwd primer	1µl
	Rev primer	1µl
	Template	0.5µl
	$MgCl_2$**	0 to 1.5µl
	DMSO*	0 to 1µl
	dNTP mix	3.5µl
	Phusion II enzyme	0.5 to 1µl

*Table 3: assembly of the PCR mixures (dNTP stock solution concentration was 2mM, primers are at 5µM, and $MgCl_2$ at 10mM Notes: *This additive was only used in cases the reaction without did not yield any product, e.g. with the DOCK5 construct. **This additive was only applied when attempting PCR from the cDNA pool, not from plasmid preps.)*

PCR reaction			
Steps (on a programmable thermostat)	**Temperature**	**Duration**	**Cycles**
	98ºC	10 min *	1
	98ºC	2 min	1
	98ºC	15 sec	25
	>(Tm+10 C) **	30 sec	
	72ºC	(L/1000) min ***	
	72ºC	10 min	1

Table 4: Protocol of PCR reactions, executed on a programmable termostat
*(Notes: *This extra pre-denaturation step was applied before the addition of dNTP and the enzyme, in case of cloning of full proteins from the cDNA pool was attempted (as with AAKG2, etc.).*
***The annealing temperature depended on the estimated Tm of the primer-template complex, it was usually set between 55 C and 65 C. ***The extension time was an explicit function of the expected product length (L, in # of nucleotides).*

Protein expression and purification

All proteins used in the current studies were produced in the Rosetta *E. coli* strain, with the standard Lac-operon based expression system using isopropyl beta-D-thiogalactoside (IPTG) as inducer. Cells were grown in appropriate antibiotics-containing LB media until they have reached an OD between 0.3 and 0.6 Unless otherwise indicated, all constructs were then induced by addition of 0.1mM IPTG, expressed at 25ºC for 4 hours, and then centrifuged (~4000 RPM, 10 min), re-suspended in PBS and centrifuged again, then flash-frozen on liquid nitrogen. Pellets that were not used immediately, were stored at -80ºC until processed.

Lysis of cells was done by the addition of a lysis buffer for Ni-NTA chromatography onto the pellets (containing 300mM NaCl, 50mM phosphate buffer [pH=8.0], 20mM imidazole, 2mM beta-mercapto-ethanol and either 0.1% IGEPAL or 0.2% CHAPS as detergent). Once lysis was complete (with the help of vigorous stirring and sonication), the lysate was centrifuged at 20,000 RPM for 20 minutes. Depending on the expected protein yield, 0.5 to 2ml Ni-NTA resin (50% slurry, per 1 litre of original culture) was added to the supernatant. The binding was aided by placing the tube containing the samples into a rotator on 4ºC for 30 minutes. Afterwards, the lysates were loaded onto fast-flow columns and washed by 20 to 40 column bed volumes (pure resin), sequentially with wash buffer I (high imidazole buffer, 300mM NaCl, 20mM TRIS [pH=8.0], 40mM imidazole and 2mM beta-mercapto-ethanol) and wash buffer II (high salt buffer, 1000mM NaCl, 20mM TRIS [pH=8.0], 20mM imidazole and 2mM beta-mercapto-ethanol). Finally, proteins were eluted with a Ni-NTA elution buffer (200mM NaCl, 20mM TRIS [pH=8.0], 400mM imidazole, 10% glycerol, 2mM beta-mercapto-ethanol and 0.2% BOG as detergent) For each sample, the reducing agent tricarboxy-ethyl-phosphine (TCEP) [pH=7.0] was added at 2mM concentration immediately after elution. Approximate yields were calculated through a Bradford assay (adding 2 to 5 µl of the eluted sample to 1ml Bradford reagent and recording the absorption at λ=595nm on a spectrophotometer against a blank sample; then concentration could be estimated using a calibration curve based on BSA). Purity of the eluted proteins was assessed by SDS-PAGE and Coomassie staining of polyacrylamide gels. The final protein solutions were flash-frozen on liquid nitrogen and stored at -80ºC.

For an additional affinity purification step (when necessary), we used either GST-sepharose or maltose resin, depending on the purification tag present in the construct (GST: glutathione-sulphuryltransferase or MBP: maltose binding protein). The protocol of these two purification methods were similar. In each case, the Ni-NTA eluted samples were diluted at least 3 times using a dilution buffer (50mM NaCl,

20mM TRIS [pH=8.0], 2mM beta-mercapto-ethanol) to reduce salt concentrations which could intefere with GST or MBP binding. At the next step, GST (or MBP) resin was added (quantity was based on the amount of estimated total protein; usually ~1ml GST resin [50%] per 5mg protein). Binding was done on a rotator similarly to Ni-NTA; the samples were loaded onto a fast-flow column and washed with 10-20 column bed volumes of wash buffer (300mM NaCl, 20mM TRIS, 20mM TRIS [pH=8.0], 2mM beta-mercapto-ethanol). The elution was done over a period of 10 minutes while gently re-suspending the resin multiple times. The elution buffer depended on the purification tag used. (The GST elution buffer contained 100mM NaCl, 20mM TRIS [pH=8.0], 10% glycerol, 02.% BOG, 2mM beta-mercapto-ethanol, and 10mM glutathion as eluent. In the case of MBP, a similar buffer was used that had 10mM maltose as eluent.)

When purifying proteins with large, intrinsically disordered segments, bacterial proteases present in the primary lysate (as well as at trace amounts in purified fractions) represented another problem.
To prevent degradation of unstructured protein segments, these lysates were treated with protease inhibitor cocktails (cOmplete EDTA-free inhibitor cocktail tablet, Roche), phenylmethanesuphonyl fluoride (PMSF) and benzamidine. All buffers following Ni-affinity purification contained PMSF (0.4 mM), benzamidine (2mM) and EDTA (1mM). This method was specifically used for production of high quality artificial substrates for the purpose of solid-phase phosphorylation experiments.

All dephosphorylated and activated MAPKs were purified (after the initial Ni-NTA capture step) through ion exchange chromatography. This was done mostly on either resource-Q or resource-S columns depending on the estimated pI of the protein. Proteins were dialized overnight against a buffer corresponding to the conditions of the chromatography (e.g. 50mM NaCl, 2mM DTT and 20mM HEPES [pH=7.0] for the S column and 50mM NaCl, 2mM DTT and 20mM TRIS (pH=8.0) for the Q column). During dialysis, we used a buffer with a volume atleast 50 times larger than the sample volume (loaded into a semi-permeable dialysis sac, usually with a 12-14 kDa cutoff range). For the purpose of chromatography, two buffers were used, buffer “A” identical to the dialysis buffer, and buffer “B”, only differing in the salt content, that was set to 1M NaCl. The purification process took place on an Äkta explorer device, with the gradient set to 0-100% over 30 minutes on default. Samples were taken at the volume 0.5-1ml; only the samples belonging to the same, clearly separable peak were pooled afterwards. Typical purities obtained after ion exchange chromatography (as assessed by SDS-PAGE) were over 95%.

Protein immobilization and solid-phase phosphorylation assays

Dot-blot experiments were carried out on synthetic substrates, purified by Ni-NTA and GST affinity chromatography in the presence of protease inhibitors and checked for purity and degradation products by SDS-PAGE after the last step. Since detergents were found to influence immobilization efficiency, care was taken to use the same detergent, BOG (Octyl-β-D-glucopyranozide), in all final solutions (including dilutions), at the same concentration (2 mg/l). Stocks containing the purified artificial substates were diluted to equal concentrations (~1mg/ml), and printed (1µl) as triplicates on nitrocellulose membranes using a Hamilton pipetting robot (GE healthcare), with the help of László Végner. Membranes were dried thoroughly afterwards on room temperature for at least 1 hour. Prior to phosphorylation, dry membranes were blocked in Tris-buffered saline and Tween 20 buffer (TBS-T) containing 3% Bovine Serum Albumin (BSA) for 30 minutes, and washed 3 times with TBS-T.

Phosphorylation was performed in volumes 5 to 10 ml, in a kinase buffer containing 50mM TRIS-HCl (pH=7.5), 10mM $MgCl_2$, 2mM DTT, 0.1% BSA and 2 mM ATP. Activated MAPKs were applied in 100-300nM concentration. The phosphorylation solution was always gently pre-mixed for at least 1 minute before application to avoid inhomogenities on the membranes. The reaction took place at room temperature on a rocker, and was stopped after 10 minutes by the addition of EDTA (at 25mM end concentration) and subsequently washed 3 times with TBS-T. Thereafter, membranes were blocked again by 3%BSA in TBS-T and developed by standard western blot techniques using an anti-phospho-T71 ATF2 antibody (Cell Signaling Technology, #9221S) at 1:1000 dilution and a secondary anti-rabbit antibody (Cell Signaling Technology, #7074S) at 1:2000 dilution. After development with the Immobilion ECL kit (Millipore), phosphorylation signal was read either by luminescence (Alpha Innotech gel documentation system) or by fluorescence (Typhoon Trio+ scanner, GE). Non-phosphorylated membranes were also checked for protein immobilization efficiency. The C-terminal hexa-histidine epitopes of GST phosphorylation reporter constructs were detected by a standard anti-His6 western blot. Dot-blot experiments were performed for each construct at least twice, using different protein stocks. Only those constructs, that consistently performed in all experiments above the non-cognate control, were regarded as “positives”.

Pull-down experiments

Aliquots of Ni-NTA purified proteins with sufficient concentration and purity were first diluted by a factor of 3, by the addition of a "dilution buffer" (20mM TRIS pH=8.0, 2mM beta-mercapto-ethanol) to negate the effect of high salt elution buffers from previous purification steps. A sufficient quantity of pre-equilibrated glutathion-sepharose or maltose resin (with estimated binding capacity below the total quantity of protein) was added subsequently, and it was allowed to bind the GST- or MBP-tagged bait on a rotator for at least 30 minutes. Then the suspension was filled onto fast-flow columns, washed (with 1-1 ml of the same buffer used for GST purification), transferred into separate tubes, centrifuged, and had the supernatant removed and exchanged for binding buffer (50 mM NaCl, 20mM TRIS pH=8.0, 2mM beta-mercapto-ethanol, 0.1% IGEPAL) to yield approximately 50% slurries upon mixing. Loading of different bait proteins was checked by running samples of each resin stock on SDS-PAGE and subsequent Coomassie staining. For each experiment, 5-20 µl resin (10-40µl slurry) was pipetted into separate tubes, containing binding buffer and equal prey quantities (to a final volume of 200µl) In case of unequal binding of baits, resin quantities were adjusted accordingly. Prey proteins were typically applied at a 50-100µg total quantity per each tube. Binding was always done at room temperature, over 30 minutes, with gentle stirring after the 15 minute mark. Then the tubes were centriguged, the supernatant was rapidly removed and exchanged for wash buffer (100mM NaCl, 20mM TRIS pH=8.0, 2mM beta-mercapto-ethanol and the detergent 0.1% IGEPAL), repeated three times over. After the final step, SDS-loading mix was added to the resin (equal to the resin volume), the samples were denatured (98ºC / 1 minute) and analysed on SDS-PAGE.

Testing of synthetic D-motif peptide arrays

The synthesis of 12 peptides (each of them 18 amino acids long) as well as their coupling to a cellulose-based hydrophilic support was ordered from IntavisAG, Germany. The glass slides (each containing two identical replicate series) were stored in an intact package on 4ºC until used. The peptide sequences are shown on the following table.

Spot #	Type of control	Peptide name	Peptide sequence
1	Standard for JNK1	pepJIP1	DTYRPKRPTTLNLFPQVP
2	Standard for JNK1	pepNFAT4	LERPSRDHLYLPLEPSYR
3	Epitope exposure control	pepNFAT4_right_shifted	ERPSRDHLYLPLEPSYRE
4	Hydrophobicity control	pepMcl1	PEPLGKRPAVLPLLELVG
5	Hydrophobicity control	pepPDE4B	EGDGISRPTTLPLTTLPS
6	Test for JNK1 (Pro-Pro)	pepPERK	LSISPPRPTTLSLDLTKN
7	Test for JNK1 (Pro-Pro)	pepAKAP6	VDPPDRSKLSLVLQSSYP
8	Standard for ERK2/p38α	pepMEF2A	GMNSRKPDLRVVIPPSSK
9	Standard for ERK2/p38α	pepRHDF1	LQRKKPPWLKLDIPSAVP
10	Epitope exposure control	pepRHDF1_left_shifted	SLQRKKPPWLKLDIPSAV
11	Test for p38α (Pro-Pro-Pro)	pepRHDF2	LQSRKPPNLSITIPPPEK
12	Epitope test for ERK2/p38α	pepCblB_truncated	LAQRRKPQPDPLQIPHLS

Table 5: List of peptides used in the Celluspot test array

To obtain quantifiable signals, we had to perform a binding step and standard western-blot procedure on each slide twice. First, the glass slides were blocked in TBS-T buffer containing 3%BSA over 30 minutes. For binding, I applied 70μg of JNK1 or 100 μg of p38α onto the membrane in either separate binding buffers (100mM NaCl, 20mM TRIS pH=8.0, 2mM beta-mercapto-ethanol, 2mM DTT) or simply TBS-T containing 3% BSA. (This caused no difference in the final outcome.) Binding was allowed to proceed for at least 1 hour. After removal of the binding mixture and washing the membrane 3 times with TBS-T, the primary antibody (anti-His6, Sigma) was applied to the membrane in 1:1000 dilution in TBS-T+3%BSA, for 1 hour. (However, direct application of the primary antibody into the binding mixture where MAPKs are present, yielded identical results.) Then the slides were washed 3 times with 10 ml TBS-T buffer, and the secondary antibody (anti-mouse IgG) was applied in 1:5000 dilution in TBS-T. After 1 hour incubation, and 3 further washing steps with TBS-T, the membranes were washed with PBS. At this point, usually no signal is detectable, thus the whole procedure needs to be repeated on the same slide (starting from the re-application of the MAPK in identical quantities, the primary and secondary antibodies with all washing steps). After the second western blot completed on the same glass slide, signals were readily detectable with the addition of 500ul ECL reagent mix (250ul luminol+250ul peroxide solution), on an Alpha Innotech gel documentation system, using chemiluminescence mode.

Fluorescence polarization measurements

Affinities of D-motifs were estimated by performing a fluorescence polarization titrations. This method relies on the difference in the anisotropy of emitted fluorescence, depending on the tumbling speed of molecules. Since the latter depends on the size of the molecules, small, fluorescently labelled peptides increase their fluorescence anisotropy once they bind to a much larger macromolecule. These recordings were done on a Biotek Synergy 4 and Biotek Cytation 3 plate readers, on a black-walled 348-well plate (after making the sequential dilutions in 1:2 steps on a 48-well plate)

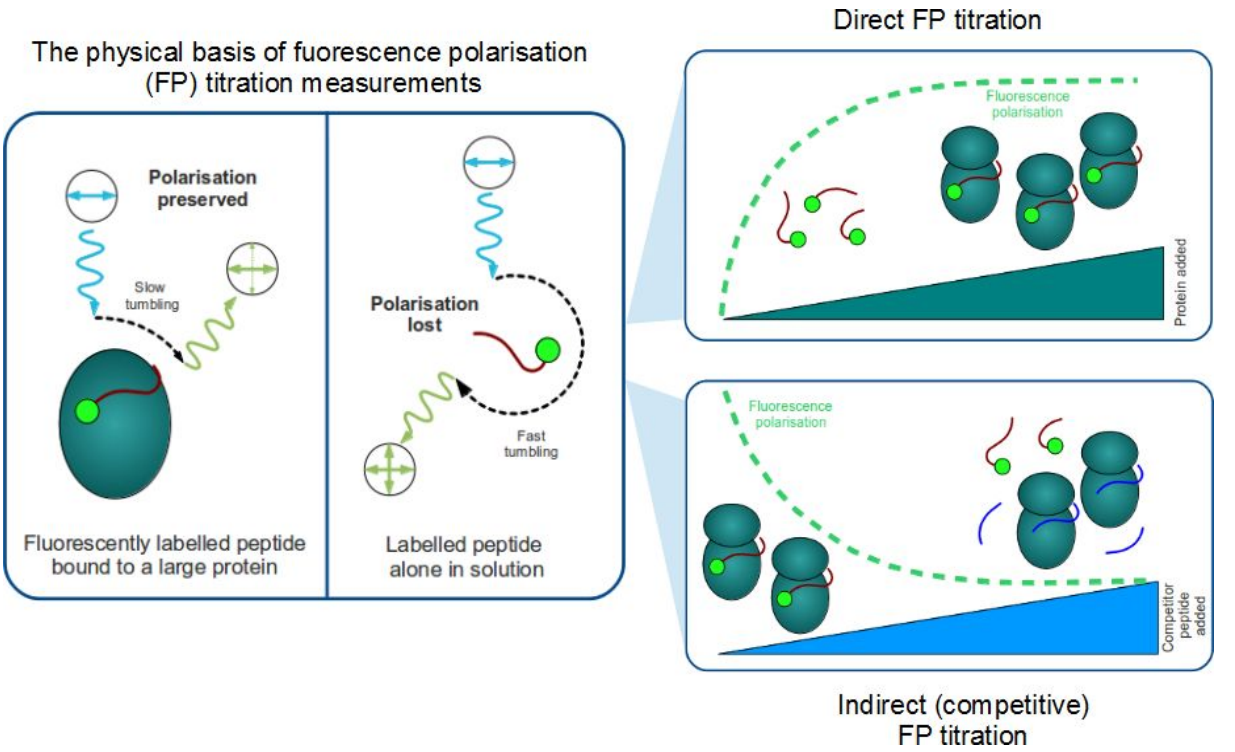


Figure 56: The basics of fluorescence polarization (FP) measurements, direct and indirect FP titrations.

Titrating a labelled peptide against increasing protein concentrations (direct FP titration) could be used to obtain the dissociation constant (Kd) value of the labelled control peptide. At the next step, the complex was titrated against unlabelled test peptides (representing novel D-motifs), starting from an appropriate concentration (typically, 80% of the saturating concentration in direct titrations). The latter experiments (indirect titrations) were utilized to determine the dissociation constants for all novel D-

motif representing peptides. Several control peptides were utilized: most were labelled with carboxyfluorescein (CF). For this case, we used the green filter cube: 485 nm excitation and 528 nm emission filter for recording. In a small number of cases, we also utilized peptides with a tetrametylrhodamine (TAMRA) label. In this case, a different set of filters (530nm excitation / 590nm emission) were used. All the curves obtained through anisotropy recordings were fitted with the Origin (version 7.7) software, unsing the equations as written below for direct and indirect titrations (Usually, between 100 and 1000 interation cycles were done until convergence of the iterative fitting was assessed).

Direct titration was fitted according to the following formula:

$$I = Imin + (Imax - Imin) * \frac{T_0 + K_0 + K_d - \sqrt{(T_0 + K_0 + K_d)^2 - 4T_0K_0)}}{2K_0}$$

Where $\mathbf{I_{min}}$ is the minimal anisotropy, $\mathbf{I_{max}}$ is the (asymptotically estimated) maximal intensity. $\mathbf{T_0}$ is the concentration of the protein, while $\mathbf{K_0}$ is the (fixed) concentration of the labelled peptide and $\mathbf{K_d}$ is the dissociation constant. In case the error of fitting was too high, $\mathbf{I_{max}}$ or $\mathbf{K_0}$ were allowed to be changed as a variable, as long as the iterative fitting of these parameters converged to an experimentally meaningful value.

Competitive titration curves were fitted with a different formula (with A B and C expressed below); The + and - signs refer to two different scenarios ($K_{dT}>K_{dB}$ or $K_{dT}<K_{dB}$):

$$I = I_{min} + (I_{max} - I_{min}) * \left\{ -\frac{\sqrt{\mathbf{A}}}{3B_0} \left[\cos\left(\frac{1}{3}\arccos(\mathbf{B})\right) \pm \sqrt{3}\sin\left(\frac{1}{3}\arccos(\mathbf{B})\right) \right] - \frac{\mathbf{C}}{3B_0} \right\}$$

$$\mathbf{A} = \mathbf{C}^2 - 3B_0D_0 + \frac{3B_0K_{dT}(B_0 + D_0 + K_{dB}) + 3B_0T_0K_{dB}}{K_{dB} - K_{dT}}$$

$$\mathbf{B} = \frac{9\mathbf{C}\left[B_0D_0 - \frac{B_0K_{dT}(B_0+D_0+K_{dB})+B_0T_0K_{dB}}{K_{dB}-K_{dT}}\right] - 2\mathbf{C}^3 - \frac{27B_0^2D_0K_{dT}}{K_{dB}-K_{dT}}}{2\sqrt{-\left\{3B_0D_0 - \frac{3B_0K_{dT}(B_0+D_0+K_{dB})+3B_0T_0K_{dB}}{K_{dB}-K_{dT}} - \mathbf{C}^2\right\}^3}}$$

$$\mathbf{C} = \frac{T_0K_{dB} + B_0K_{dT}}{K_{dB} - K_{dT}} - (B_0 + D_0 + K_{dB})$$

Here $\mathbf{I_{min}}$ is the minimal anisotropy, $\mathbf{I_{max}}$ is the (asymptotically estimated) maximal anisotropy. $\mathbf{B_0}$ is the concentration of the labelled peptide, $\mathbf{D_0}$ is the (constant) concentration of the protein (corresponding to 80% saturation of the labelled peptide-protein complex), and $\mathbf{K_{dB}}$ is the dissociation constant measured by the direct titrations, while $\mathbf{K_{dT}}$ is the dissociation constant of the unlabelled competitor peptide. Similarly to direct titrations, in case of divergence or grossly erroneous fits, select variables (I_{max}, K_{dT}, D_0) were allowed to vary duting the fitting iterations, as long as they converged to yield sensible values. In all cases, the direct titrations were done the same day, using the same protein batch as for the indirect titrations to minimize the error of FP titrations.

	MAPK partner		
Unlabelled competitor	ERK2	JNK1	p38α
AAKG2	CF-HePTP	TAMRA-NFAT4	TAMRA-MEF2A
CCNT2	CF-HePTP	TAMRA-NFAT4	TAMRA-MEF2A
IRS1	CF-HePTP	TAMRA-NFAT4	TAMRA-MEF2A
DOCK5	CF-HePTP	CF-JIP3	TAMRA-MEF2A
JAZF1	CF-HePTP	CF-JIP3	TAMRA-MEF2A
MYO9B	CF-HePTP	CF-JIP3	TAMRA-MEF2A
APBA2	CF-RSK3	CF-JIP3	CF-RSK3
KSR2	CF-RSK1	TAMRA-NFAT4	CF-RSK3
ATF7	CF-RHDF1	CF-JIP3	CF-RHDF1
DCX	CF-RHDF1	CF-JIP3	CF-RHDF1
DOCK7	CF-RHDF1	CF-JIP3	CF-RHDF1
GAB3	CF-RHDF1	CF-JIP3	CF-RHDF1
GAB3+	CF-RHDF1	CF-JIP3	CF-RHDF1
MKP5	CF-RHDF1	CF-JIP3	CF-RHDF1
PDE4B	CF-RHDF1	CF-JIP3	CF-RHDF1
RHDF1	CF-RHDF1	CF-JIP3	CF-RHDF1

Table 6: list of competitor peptides used for each titration mentioned in the current study

The sequences of the **competitor peptides** were the following: HePTP: RLQERRGSNVALML, NFAT4: LERPSRDHLYLPLE, MEF2A: SRKPDLRVVIPPSS, JIP3: RKERPTSLNNFPL, RSK3: PVLEPVGRSTLAQRRGIKKITSTAL, RSK1: PQLKPIESSILAQRRVRKLPSTTL, RHDF1: SLQRKKPPWLKLDIPS. In all cases, the label was linked to the N-terminus of the peptides, by direct coupling of the dye to the -NH_2-group with the same Fmoc strategy as the peptide was synthesized.

Bimolecular fluorescent complementation assays (BiFC)

The full length cDNA of yellow fluorescent protein (YFP) was split at amino acid residue 159 and fragments encoding the bigger N-terminal (F1) and the smaller C-terminal (F2) half were pasted into pcDNA 3.1 vectors (Invitrogen). ERK2 was expressed as C-terminal, p38α and JNK1 as N-terminal F2 fusions. However, we noted that constructs with wild-type JNK1 and p38α yielded poor expression, possibly due to their deleterious effects on cell survival. To facilitate expression, JNK1 and p38α had kinase inactivating mutations (K55R and K53R, respectively), while ERK2 was wild-type. The ERK2 and JNK1 constructs contained a FLAG tag, but similarly tagged p38α constructs could not be expressed to a comparable extent. Therefore expression levels of F2-p38α could only be monitored by an anti-p38α antibody. MAPK partners were expressed as N-terminal and C-terminal F1 fusions with FLAG tags. F1 and F2 fusion pairs that gave the highest BiFc signal with wild-type MAPK partners were chosen to analyze the impact of docking motif truncations or mutations. These were introduced into full length MAPK partner constructs by PCR or by the QuickChange method. All sequences were verified by DNA sequencing.

We noted that complementation efficiencies of different fusion constructs varied wildly, with prominent steric effects. The highest signals were often obtained with test proteins where the fusion tag was as close to the docking motif as possible (with an N-terminal tag in case of proteins where the docking motif is located N-terminally and with a C-terminal tag in proteins where the D-motif is close to the natural C-terminus of the protein). Therefore this setup was chosen as our "default" for most novel partner proteins.

For the purpose of transfection, HEK293T cells were cultured in Dulbecco’s modified Eagle’s medium (DMEM, Lonza) containing 10% fetal bovine serum and 1% penicillin/streptomycin at 37 °C in an atmosphere of 5% CO2 in 25 cm^2 tissue culture flasks (Orange Scientific). Cells were seeded onto 48-well plate (tissue culture test plate 96F, TPP) at 60-70% confluence 24 hrs prior to transfection. The medium was then changed to serum reduced OPTI-MEM (Gibco). Transient transfections with Lipofectamine 2000 reagent (Invitrogen) were carried out according to the manufacturer’s instruction. Cells were assayed 2 days post-transfection. For BiFc signal intensity measurements cells were washed and suspended in 100 µl PBS. 20 µl of this cell suspension (~20,000 cells) was aliquoted into a 384-well black-sided plate.

Fluorescence intensity per well was measured using a Synergy H4 (BioTek Instruments) fluorescence plate-reader (excitation/emission wavelength was 515/535 nm). Subsequently, 50µl samples of pooled cells resuspended in PBS were collected and subjected to Western-blots, using anti-FLAG tag antibody (Sigma, F1804) and anti- p38α antibody. For imaging, transfected cells were examined with an Olympus IX81 microscope using an Olympus FluoView 500 confocal laser scanning microscope system (Hamburg, Germany). YFP fluorescence was imaged using 514 nm excitation and a 535-560 nm emission filter.

PSSM building, sequence logos and final scoring

Position specific-scoring matrices for JIP1, NFAT4, greater MEF2A, and greater DCC classes were built including already known and recently validated human motifs as well as all their identifiable vertebrate orthologs. This was mainly done by Tomas Bastys (MPI, Saarbrücken). In such a matrix each row represents one of 20 possible residues, and each column represents a position in a motif. Thus the score for residue X at position i is defined in the following way:

$$X_i = \frac{\sum_s w_s * Ind(s_i{=}X) + P * X_b}{\sum_s w_s + P}$$

where ***s*** is a peptide, $\mathbf{w_s}$ is the weight of that peptide based on the species from which it stems, **Ind** is the indicator function (which is 1 when its argument is true and 0 otherwise), **P** is the pseudo-count defined as square-root of total number of training peptides from the class (used to account for residues that don't appear at position i), and $\mathbf{X_b}$ is the background frequency of the residue (based on Uniprot release 2013.5) For computational efficiency and to account for background frequencies of residues, log-odds scores of $\mathbf{X_i}$ were used in the form $\mathbf{X_i} = \mathbf{log}\ (\mathbf{X_i} / \mathbf{X_b})$. The final score was calculated as the sum of the log-odds scores of individual positions. The distance between species A and B was the direct sum of branch lengths leading from A endpoint to B endpoint along the tree. These distances were the same that appear in PhyloWidget visualization of Ensembl Compara gene trees.[192] In case of species not included there, evolutionary distances were extracted from trees published in literature and numerically re-calculated to fit the scaling of Ensembl-derived distances, wherever possible. In the remaining few species with only the tree topology, but no exact distance metrics available, a numeric interpolation with equal weights was used.

Sequence logos were constructed using Seq2Logo-2.0. Height of residue X at position i is directly proportional to it's PSSM score $\mathbf{X_i}$ (with p = 0) and $\mathbf{R_i}$, information content of position i, defined as:

$$R_i = log_2(20) - (H_i + e_n) \qquad H_i = -\sum X_i * log_2(X_i)$$

where $\mathbf{e_n}$ is the small sample correction parameter, expressed as 19 / (2 * ln(2) * number of peptides), and $\mathbf{H_i}$ is uncertainty at position i, defined by the second equation. The receiver operating characteristic (ROC) were constructed by adding simulated negative cases: peptides in the human proteome conforming to the respective D-motif consensus but lying in a Pfam A structured region were scored together with validated true positive motifs, and at every value of the true positive rate the corresponding value of the false positive rate was calculated and plotted. The area under the ROC (AUC) curve was calculated to assess the quality of the prediction: the closer it is to 1 the better is the predictor, and 0.9 indicates a very good predictor. For each motifs class, AUC calculation was performed in 5-fold cross-validation setting with 100 samples, selecting 4 folds of D-motifs for training and 1 remaining fold, together with the D-motifs from Pfam A domains, for testing, and not allowing motifs from the same vertebrate orthology set to be used both for training and testing.

RÖVID ÖSSZEFOGLALÁS

A mitogén aktivált proteinkinázok (MAPK-k) számtalan éltettani folyamat szabályozásában töltenek be fontos szerepet. A katalitikus helyük által megkövetelt konszenzus igen laza, cserébe viszont eme kinázok a szubsztrátjaikat, illetve egyéb szabályozó fehérjéiket extra interakciós elemek, dokkoló motívumok segítségével ismerik fel. Noha a dokkoló motívumok legfontosabb típusát (az úgynevezett D-motívumokat) több mint egy évtizede ismerjük, eddig nem volt mód arra, hogy megbízhatóan megjósoljuk a jelenlétüket csupán a fehérjék szekvenciájának ismeretében. A jelen kutatási téma elsődleges célja épp egy ilyen módszer kidolgozása volt.

Először a D-motívumok korábbi definícióját kellett kijavítanunk, méghozzá a rendelkezésre álló, kísérletesen meghatározott szerekezetek és szerkezeti modellek segítségével. Ezután megalkottunk egy olyan kereső és szűrő algoritmust, ami az emberi proteomban található összes potenciális D-motívumot megtalálja. A találatok "jóságának" megállapítására homológia-modellezés alapú energiabecslést, illetve szerkvencia-alapú (pontozómátrix) módszereket használtunk. A potenciális motívumok tesztelésére egy teljesen újfajta esszét fejlesztettünk ki, amely mesterséges szubsztrátok szilárd fázison történő foszforilációján alapszik. Az ezen kísérleti módszerrel észlelt találatokat más *in vitro* módszerrel is megerősítettük, illetve bebizonyítottuk, hogy legalábbis az interakciók egy része sejten belül, teljes fehérjék esetében is megvalósul. Természetesen nem minden MAPK partner jósolható meg szisztematikus kereséssel, csak azok, amelyek "szokványos" lináris motívummal rendelkeznek.

Az általunk azonosított nagy számú, potenciálisan új MAPK partner-fehérje lehetővé tette, hogy a MAP kináz rendszer kölcsönhatásainak evolúcióját behatóbban is elemezzük. Kiderült hogy - míg a kinázok maguk nagyon ősiek és konzerváltak - a partnereik relatíve gyorsan cserélődnek: A humán proteomban megtalálható legtöbb D-motívum a csupán a gerincesek evolúciója során jött létre, méghozzá változatos módon. Mindezt feltehetően az teszi lehetővé, hogy a MAPK foszforiláció által szabályozott célhelyek túlnyomó többsége maga is lineáris motívumot alkot, és így evolúciójuk ugyanolyan gyors lehet, mint maguké a dokkoló motívumoké.

Az új partnerfehérjék túlnyomó többsége feltehetően - eddig még nem azonosított - szubsztrát. Ezek elemzése elárulja, hogy a MAP kinázok nemcsak a sejtosztódásban, illetve a "stresszre" adott válaszban játszanak szerepet, hanem rengeteg, finoman szabályozott, szövetspecifikus folyamatban is, mint az idegsejtek fejlődése, a szívizom anyagcseréjének szabályozása, illetve az inzulin-hatás.

BRIEF SUMMARY

Mitogen-activated protein kinases (MAPKs) paly a key role in regulating a diverse set of physiological processes. The consensus motif required by their catalytic site is very loose; however this is offset by their ability to recruit their substrates and miscellanous regulator proteins with the help of docking motifs. Although the most important type of MAPK docking motifs (termed "D-motifs") has already been known for more than a decade through numerous examples, they were not predictable in a reliable way, using protein sequences alone. Our main aim in the current study was to develop such a method.

First, we had to amend the defintion of D-motifs, with the help of experimentally-determined structures and structural models. Next, we developed an *in silico* search and filtering method to identify all suitable D-motif candidates from the human proteome. To assess the "goodness" of hits, we utilized homology modelling based energy estimations as well as sequence based (position-specific scoring matrix) methods. To test motif candidates experimentally, we designed an entirely new assay, based on phosphorylation of artificial substrates immobilized on a solid phase. We further validated hits from the former assay by other *in vitro* methods, and tested a select few in full protein context, in living cells in order to prove that they are also functional in a biological setting. Not all partner proteins are predictable with such systematic approaches, only those posessign a "regular" D-motif are.

The high number potentially novel MAPK partner proteins identified in our assays also enabled us to analyze the interactions in the MAPK system from an evolutionary point of view. Although the kinases constituting this pathway are very ancient and conserved, their partners tend to change fast. The majority of D-motifs found in the human proteome only emerged during vertebrate evolution, through various mechanisms. This is likely made possible because of the nature of MAPK phosphorylation target sites, constituting linear motifs themselves, and thus being able to evolve similarly fast.

Most of the novel partner proteins are probably - still unidentified - MAPK substrates. Their analysis suggests that MAPKs are not just simply regulating cell division or "stress" responses, but also participate in a number of fine-tuned, tissue-specific processes, including neuronal development, the regulation of cardiomyocyte metabolism or insulin action.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank for all involved in the current research project, namely:

- **Anita Alexa**, for her enthusiastic help in all experiments, teaching me countless new methods and tricks, and her contribution to the project by concieveing, building and successfully executing the BiFC assays.
- **Bálint Mészáros**, who built the primary search algorithm and many filters with great expertise and thorough attention.
- **Zsuzsa Dosztányi**, who helped our project many times with her expert bioinformatic knowledge as well as with new ideas.
- **Tomas Bastys**, for his always-friendly attitude, for his contribution by many scripts, algorithms and ideas, as well as realizing the PSSM-based searches.
- **Olga Kalinina**, for offering her best skills to develop a FoldX-based pipeline as well as performing the tedious automated evolutionary analyses with great perseverance
- **Ágnes Szonja Garai**, for her strong support, both experimental and otherwise, as well as for the major bulk of FP titrations and their analysis.
- **Klára Pongorné Kirsch** , for her aid, and contribution with many additional measurements.
- **László Végner**, for his technical knowledge in robotics, without which the dot-blot arrays would never have been printed.
- **Gergő Gógl**, for helping to create our first structural models.
- **Marianna Rakács**, our lab technician, for her constant, but "invisible" help and contribution to all experiments.

I would like to thank our P.I., **Attila Reményi**, who provided us his extensive research expertise, both the material and the instrumental background, coordinated the project and organized multiple collaborations so essential for our success.

All other past and present members of our lab also deserve many thanks, for their unconditional help, the lifting mood, for the many intellectual discussions and great time we had together:

- **Gábor Glatz**, our first home-grown postdoc
- **János Varga-Kugler**, ex-member and friend
- **Imre Törő**, senior ex-member
- **Ádám Pót**i, PhD student
- **Péter Sok**, PhD student
- **Orsolya Cseh** & **Lili Zsákai**, ex-members
- **Boglárka Zámbó**, **Ferenc Fördős** & **Martina Rádli**, alumni
- **Melinda Lukács** & **Tünde Bárkai**, founding ex-members
- and many others...

Last but not least, I would like to express my thanks to my parents, **Anikó Zeke** & **László Zeke** who tirelessly laboured all time to provide me the perfect family background, without whose support I could never have become a true scientist. Similarly, I must also thank my brother, **László Tamás Zeke**, who also helped me with the preparation of this manuscript.

PUBLICATIONS

Published scientific articles:

- **Scaffolds: interaction platforms for cellular signalling circuits.** (Review). Zeke A, Lukács M, Lim WA, Reményi A. *Trends in Cell Biology* 2009 Aug;19(8):364-74.

- **Specificity of linear motifs that bind to a common mitogen-activated protein kinase docking groove.** Garai Á, Zeke A, Gógl G, Törő I, Fördős F, Blankenburg H, Bárkai T, Varga J, Alexa A, Emig D, Albrecht M, Reményi A. *Science Signalling* 2012 Oct 9;5(245):ra74.

- **Structural assembly of the signaling competent ERK2-RSK1 heterodimeric protein kinase complex.** Alexa A, Gógl G, Glatz G, Garai Á, Zeke A, Varga J, Dudás E, Jeszenői N, Bodor A, Hetényi C, Reményi A. *Proceedings of the National Academy of Sciences, U S A.* 2015 Mar 3;112(9):2711-6.

Articles to be published, pertaining to the current study:

- **Mapping mitogen-activated protein kinase interactors: An extensive network with fast-evolving partnerships.** Zeke A. Bastys T, Alexa A, Mészáros B, Garai Á, Kirsch K, Kalinina O, Dosztányi Zs, Reményi A. 2015 (Under review)

Participations on international conferences with posters/presentations:

- **EMBO meeting 2011**, Vienna, Austria. (Poster): MAP kinase – linear motif interactions restrain signaling specificity in paralogous pathways. Á. Garai, A. Zeke, F. Fördős, G. Gógl, A. Reményi

- **FEBS3+ conference 2012**, Opatija, Croatia (Poster): Searching for new MAP kinase substrates with a novel *in silico* method. A. Zeke, Á. Garai, O. Kalinina, B. Mészáros, H. Blankenburg, M. Albrecht, Zs. Dosztányi, A. Reményi.

- **Interdisciplinary Signalling Workshop 2014**, Visgrád, Hungary (Presentation): Identifying novel Mitogen Activated Protein Kinase (MAPK) partners with a combination of sequence-based, structural modelling & evolutionary analyses. A. Zeke, A. Alexa, Á. Garai, O. Kalinina, T. Bastys, B. Mészáros, Zs. Dosztányi, A. Reményi

LITERATURE REFERENCES

1. Hoshi, M., Nishida, E. & Sakai, H. Activation of a Ca2+-inhibitable protein kinase that phosphorylates microtubule-associated protein 2 in vitro by growth factors, phorbol esters, and serum in quiescent cultured human fibroblasts. *J. Biol. Chem.* **263,** 5396–401 (1988).

2. Hoshi, M., Nishida, E., Inagaki, M., Gotoh, Y. & Sakai, H. Activation of a serine/threonine kinase that phosphorylates microtubule-associated protein 1B in vitro by growth factors and phorbol esters in quiescent rat fibroblastic cells. *Eur. J. Biochem.* **193,** 513–9 (1990).

3. Gotoh, Y. *et al.* Microtubule-associated-protein (MAP) kinase activated by nerve growth factor and epidermal growth factor in PC12 cells. Identity with the mitogen-activated MAP kinase of fibroblastic cells. *Eur. J. Biochem.* **193,** 661–9 (1990).

4. Hibi, M., Lin, A., Smeal, T., Minden, A. & Karin, M. Identification of an oncoprotein- and UV-responsive protein kinase that binds and potentiates the c-Jun activation domain. *Genes Dev.* **7,** 2135–48 (1993).

5. Han, J., Lee, J. D., Tobias, P. S. & Ulevitch, R. J. Endotoxin induces rapid protein tyrosine phosphorylation in 70Z/3 cells expressing CD14. *J. Biol. Chem.* **268,** 25009–14 (1993).

6. Han, J., Lee, J. D., Bibbs, L. & Ulevitch, R. J. A MAP kinase targeted by endotoxin and hyperosmolarity in mammalian cells. *Science* **265,** 808–11 (1994).

7. Jiang, Y. *et al.* Characterization of the structure and function of a new mitogen-activated protein kinase (p38beta). *J. Biol. Chem.* **271,** 17920–6 (1996).

8. Boulton, T. G. *et al.* ERKs: a family of protein-serine/threonine kinases that are activated and tyrosine phosphorylated in response to insulin and NGF. *Cell* **65,** 663–75 (1991).

9. Lechner, C., Zahalka, M. A., Giot, J. F., Møller, N. P. & Ullrich, A. ERK6, a mitogen-activated protein kinase involved in C2C12 myoblast differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 4355–9 (1996).

10. Abe, M. K. *et al.* ERK8, a new member of the mitogen-activated protein kinase family. *J. Biol. Chem.* **277,** 16733–43 (2002).

11. Widmann, C., Gibson, S., Jarpe, M. B. & Johnson, G. L. Mitogen-activated protein kinase: conservation of a three-kinase module from yeast to human. *Physiol. Rev.* **79,** 143–80 (1999).

12. Xia, Z., Dickens, M., Raingeaud, J., Davis, R. J. & Greenberg, M. E. Opposing effects of ERK and JNK-p38 MAP kinases on apoptosis. *Science* **270,** 1326–31 (1995).

13. Matsuda, S., Kawasaki, H., Moriguchi, T., Gotoh, Y. & Nishida, E. Activation of protein kinase cascades by osmotic shock. *J. Biol. Chem.* **270,** 12781–6 (1995).

14. Clerk, A., Fuller, S. J., Michael, A. & Sugden, P. H. Stimulation of ‘stress-regulated’ mitogen-activated protein kinases (stress-activated protein kinases/c-Jun N-terminal kinases and p38-mitogen-activated protein kinases) in perfused rat hearts by oxidative and other stresses. *J. Biol. Chem.* **273,** 7228–34 (1998).

15. Lee, J. C. *et al.* A protein kinase involved in the regulation of inflammatory cytokine biosynthesis. *Nature* **372,** 739–46

16. Segall, J. E. *et al.* A MAP kinase necessary for receptor-mediated activation of adenylyl cyclase in Dictyostelium. *J. Cell Biol.* **128,** 405–13 (1995).

17. Kosetsu, K. *et al.* The MAP kinase MPK4 is required for cytokinesis in Arabidopsis thaliana. *Plant Cell* **22,** 3778–90 (2010).

18. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298,** 1912–34 (2002).

19. Lavoie, H. & Therrien, M. Regulation of RAF protein kinases in ERK signalling. *Nat. Rev. Mol. Cell Biol.* **16,** 281–298 (2015).

20. Rajakulendran, T., Sahmi, M., Lefrançois, M., Sicheri, F. & Therrien, M. A dimerization-dependent mechanism drives RAF catalytic activation. *Nature* **461,** 542–5 (2009).

21. Emery, C. M. *et al.* MEK1 mutations confer resistance to MEK and B-RAF inhibition. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 20411–6 (2009).

22. Macdonald, S. G. *et al.* Reconstitution of the Raf-1-MEK-ERK signal transduction pathway in vitro. *Mol. Cell. Biol.* **13,** 6615–20 (1993).

23. McKay, M. M., Ritt, D. A. & Morrison, D. K. Signaling dynamics of the KSR1 scaffold complex. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 11022–7 (2009).

24. Samatar, A. A. & Poulikakos, P. I. Targeting RAS–ERK signalling in cancer: promises and challenges. *Nat. Rev. Drug Discov.* **13,** 928–942 (2014).

25. Takekawa, M. & Saito, H. A family of stress-inducible GADD45-like proteins mediate activation of the stress-responsive MTK1/MEKK4 MAPKKK. *Cell* **95,** 521–30 (1998).

26. Saitoh, M. *et al.* Mammalian thioredoxin is a direct inhibitor of apoptosis signal-regulating kinase (ASK) 1. *EMBO J.* **17,** 2596–606 (1998).

27. Adhikari, A., Xu, M. & Chen, Z. J. Ubiquitin-mediated activation of TAK1 and IKK. *Oncogene* **26,** 3214–26 (2007).

28. Du, Y., Böck, B. C., Schachter, K. A., Chao, M. & Gallo, K. A. Cdc42 induces activation loop phosphorylation and membrane targeting of mixed lineage kinase 3. *J. Biol. Chem.* **280,** 42984–93 (2005).

29. Remy, G. *et al.* Differential activation of p38MAPK isoforms by MKK6 and MKK3. *Cell. Signal.* **22,** 660–7 (2010).

30. Brancho, D. *et al.* Mechanism of p38 MAP kinase activation in vivo. *Genes Dev.* **17,** 1969–78 (2003).

31. Garai, Á. *et al.* Specificity of linear motifs that bind to a common mitogen-activated protein kinase docking groove. *Sci. Signal.* **5,** ra74 (2012).

32. Lovett, F. A., Cosgrove, R. A., Gonzalez, I. & Pell, J. M. Essential role for p38alpha MAPK but not p38gamma MAPK in Igf2 expression and myoblast differentiation. *Endocrinology* **151,** 4368–80 (2010).

33. Eckert, R. L. *et al.* p38 Mitogen-activated protein kinases on the body surface--a function for p38 delta. *J. Invest. Dermatol.* **120,** 823–8 (2003).

34. Hasegawa, M. *et al.* Stress-activated protein kinase-3 interacts with the PDZ domain of alpha1-syntrophin. A mechanism for specific substrate recognition. *J. Biol. Chem.* **274,** 12626–31 (1999).

35. Ganiatsas, S. *et al.* SEK1 deficiency reveals mitogen-activated protein kinase cascade crossregulation and leads to abnormal hepatogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 6881–6

(1998).

36. Wada, T. *et al.* MKK7 couples stress signalling to G2/M cell-cycle progression and cellular senescence. *Nat. Cell Biol.* **6,** 215–26 (2004).

37. Yamasaki, T. *et al.* Stress-activated protein kinase MKK7 regulates axon elongation in the developing cerebral cortex. *J. Neurosci.* **31,** 16872–83 (2011).

38. Kuan, C. Y. *et al.* The Jnk1 and Jnk2 protein kinases are required for regional specific apoptosis during early brain development. *Neuron* **22,** 667–76 (1999).

39. Chao, T. H., Hayashi, M., Tapping, R. I., Kato, Y. & Lee, J. D. MEKK3 directly regulates MEK5 activity as part of the big mitogen-activated protein kinase 1 (BMK1) signaling pathway. *J. Biol. Chem.* **274,** 36035–8 (1999).

40. Nakamura, K. & Johnson, G. L. PB1 domains of MEKK2 and MEKK3 interact with the MEK5 PB1 domain for activation of the ERK5 pathway. *J. Biol. Chem.* **278,** 36989–92 (2003).

41. Faurobert, E. & Albiges-Rizo, C. Recent insights into cerebral cavernous malformations: a complex jigsaw puzzle under construction. *FEBS J.* **277,** 1084–96 (2010).

42. Yang, J. *et al.* Mekk3 is essential for early embryonic cardiovascular development. *Nat. Genet.* **24,** 309–13 (2000).

43. Zhou, Z. *et al.* The cerebral cavernous malformation pathway controls cardiac development via regulation of endocardial MEKK3 signaling and KLF expression. *Dev. Cell* **32,** 168–80 (2015).

44. Sunadome, K. *et al.* ERK5 regulates muscle cell fusion through Klf transcription factors. *Dev. Cell* **20,** 192–205 (2011).

45. Yan, L. *et al.* Knockout of ERK5 causes multiple defects in placental and embryonic development. *BMC Dev. Biol.* **3,** 11 (2003).

46. Spiering, D. *et al.* MEK5/ERK5 signaling modulates endothelial cell migration and focal contact turnover. *J. Biol. Chem.* **284,** 24972–80 (2009).

47. Li, T. *et al.* Targeted deletion of the ERK5 MAP kinase impairs neuronal differentiation, migration, and survival during adult neurogenesis in the olfactory bulb. *PLoS One* **8,** e61948 (2013).

48. Hayashi, M. *et al.* Targeted deletion of BMK1/ERK5 in adult mice perturbs vascular integrity and leads to endothelial failure. *J. Clin. Invest.* **113,** 1138–48 (2004).

49. Coulombe, P. & Meloche, S. Atypical mitogen-activated protein kinases: structure, regulation and functions. *Biochim. Biophys. Acta* **1773,** 1376–87 (2007).

50. Aberg, E. *et al.* Docking of PRAK/MK5 to the atypical MAPKs ERK3 and ERK4 defines a novel MAPK interaction motif. *J. Biol. Chem.* **284,** 19392–401 (2009).

51. Déléris, P. *et al.* Activation loop phosphorylation of ERK3/ERK4 by group I p21-activated kinases (PAKs) defines a novel PAK-ERK3/4-MAPK-activated protein kinase 5 signaling pathway. *J. Biol. Chem.* **286,** 6470–8 (2011).

52. Ishitani, T. *et al.* The TAK1-NLK-MAPK-related pathway antagonizes signalling between beta-catenin and transcription factor TCF. *Nature* **399,** 798–802 (1999).

53. Ishitani, S., Inaba, K., Matsumoto, K. & Ishitani, T. Homodimerization of Nemo-like kinase is essential for activation and nuclear localization. *Mol. Biol. Cell* **22,** 266–77 (2011).

54. Kortenjann, M. *et al.* Abnormal bone marrow stroma in mice deficient for nemo-like kinase, Nlk.

Eur. J. Immunol. **31,** 3580–7 (2001).

55. Bardwell, L. A walk-through of the yeast mating pheromone response pathway. *Peptides* **26,** 339–50 (2005).

56. Elion, E. A. The Ste5p scaffold. *J. Cell Sci.* **114,** 3967–78 (2001).

57. Zalatan, J. G., Coyle, S. M., Rajan, S., Sidhu, S. S. & Lim, W. A. Conformational control of the Ste5 scaffold protein insulates against MAP kinase misactivation. *Science* **337,** 1218–22 (2012).

58. Good, M., Tang, G., Singleton, J., Reményi, A. & Lim, W. A. The Ste5 scaffold directs mating signaling by catalytically unlocking the Fus3 MAP kinase for activation. *Cell* **136,** 1085–97 (2009).

59. Liu, H., Styles, C. A. & Fink, G. R. Elements of the yeast pheromone response pathway required for filamentous growth of diploids. *Science* **262,** 1741–4 (1993).

60. Cullen, P. J. & Sprague, G. F. The regulation of filamentous growth in yeast. *Genetics* **190,** 23–49 (2012).

61. Cook, J. G., Bardwell, L., Kron, S. J. & Thorner, J. Two novel targets of the MAP kinase Kss1 are negative regulators of invasive growth in the yeast Saccharomyces cerevisiae. *Genes Dev.* **10,** 2831–48 (1996).

62. Neiman, A. M. *et al.* Functional homology of protein kinases required for sexual differentiation in Schizosaccharomyces pombe and Saccharomyces cerevisiae suggests a conserved signal transduction module in eukaryotic organisms. *Mol. Biol. Cell* **4,** 107–20 (1993).

63. Dupré, A., Haccard, O. & Jessus, C. Mos in the oocyte: how to use MAPK independently of growth factors and transcription to control meiotic divisions. *J. Signal Transduct.* **2011,** 350412 (2011).

64. Saito, H. & Posas, F. Response to hyperosmotic stress. *Genetics* **192,** 289–318 (2012).

65. Maeda, T., Wurgler-Murphy, S. M. & Saito, H. A two-component system that regulates an osmosensing MAP kinase cascade in yeast. *Nature* **369,** 242–5 (1994).

66. Degols, G., Shiozaki, K. & Russell, P. Activation and regulation of the Spc1 stress-activated protein kinase in Schizosaccharomyces pombe. *Mol. Cell. Biol.* **16,** 2870–7 (1996).

67. Caffrey, D. R., O'Neill, L. A. & Shields, D. C. The evolution of the MAP kinase pathways: coduplication of interacting proteins leads to new signaling cascades. *J. Mol. Evol.* **49,** 567–82 (1999).

68. Levin, D. E. Cell wall integrity signaling in Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.* **69,** 262–91 (2005).

69. Kim, K.-Y., Truman, A. W. & Levin, D. E. Yeast Mpk1 mitogen-activated protein kinase activates transcription through Swi4/Swi6 by a noncatalytic mechanism that requires upstream signal. *Mol. Cell. Biol.* **28,** 2579–89 (2008).

70. Whinston, E., Omerza, G., Singh, A., Tio, C. W. & Winter, E. Activation of the Smk1 mitogen-activated protein kinase by developmentally regulated autophosphorylation. *Mol. Cell. Biol.* **33,** 688–700 (2013).

71. Mendoza, M. C. *et al.* Loss of SMEK, a novel, conserved protein, suppresses MEK1 null cell polarity, chemotaxis, and gene expression defects. *Mol. Cell. Biol.* **25,** 7839–53 (2005).

72. Ellis, J. G., Davila, M. & Chakrabarti, R. Potential involvement of extracellular signal-regulated kinase 1 and 2 in encystation of a primitive eukaryote, Giardia lamblia. Stage-specific activation

and intracellular localization. *J. Biol. Chem.* **278,** 1936–45 (2003).

73. Dorin, D. *et al.* An atypical mitogen-activated protein kinase (MAPK) homologue expressed in gametocytes of the human malaria parasite Plasmodium falciparum. Identification of a MAPK signature. *J. Biol. Chem.* **274,** 29912–20 (1999).

74. Dorin-Semblat, D. *et al.* Functional characterization of both MAP kinases of the human malaria parasite Plasmodium falciparum by reverse genetics. *Mol. Microbiol.* **65,** 1170–80 (2007).

75. Nagy, S. K. *et al.* Activation of AtMPK9 through autophosphorylation that makes it independent of the canonical MAPK cascades. *Biochem. J.* **467,** 167–75 (2015).

76. Rodriguez, M. C. S., Petersen, M. & Mundy, J. Mitogen-activated protein kinase signaling in plants. *Annu. Rev. Plant Biol.* **61,** 621–49 (2010).

77. Brodersen, P. *et al.* Arabidopsis MAP kinase 4 regulates salicylic acid- and jasmonic acid/ethylene-dependent responses via EDS1 and PAD4. *Plant J.* **47,** 532–46 (2006).

78. Ishihama, N. & Yoshioka, H. Post-translational regulation of WRKY transcription factors in plant immunity. *Curr. Opin. Plant Biol.* **15,** 431–7 (2012).

79. Lochhead, P. A., Sibbet, G., Morrice, N. & Cleghon, V. Activation-loop autophosphorylation is mediated by a novel transitional intermediate form of DYRKs. *Cell* **121,** 925–36 (2005).

80. Howard, C. J. *et al.* Ancestral resurrection reveals evolutionary mechanisms of kinase plasticity. *Elife* **3,** (2014).

81. Kannan, N. & Neuwald, A. F. Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2alpha. *Protein Sci.* **13,** 2059–77 (2004).

82. Zhu, G. *et al.* Exceptional disfavor for proline at the P + 1 position among AGC and CAMK kinases establishes reciprocal specificity between them and the proline-directed kinases. *J. Biol. Chem.* **280,** 10743–8 (2005).

83. Soundararajan, M. *et al.* Structures of Down syndrome kinases, DYRKs, reveal mechanisms of kinase activation and substrate recognition. *Structure* **21,** 986–96 (2013).

84. Marin, O., Meggio, F., Draetta, G. & Pinna, L. A. The consensus sequences for cdc2 kinase and for casein kinase-2 are mutually incompatible. A study with peptides derived from the beta-subunit of casein kinase-2. *FEBS Lett.* **301,** 111–4 (1992).

85. Himpel, S. *et al.* Specificity determinants of substrate recognition by the protein kinase DYRK1A. *J. Biol. Chem.* **275,** 2431–8 (2000).

86. Canagarajah, B. J., Khokhlatchev, A., Cobb, M. H. & Goldsmith, E. J. Activation mechanism of the MAP kinase ERK2 by dual phosphorylation. *Cell* **90,** 859–69 (1997).

87. Alexa, A. *et al.* Structural assembly of the signaling competent ERK2–RSK1 heterodimeric protein kinase complex. *Proc. Natl. Acad. Sci.* **112,** 201417571 (2015).

88. Emrick, M. A. *et al.* The gatekeeper residue controls autoactivation of ERK2 via a pathway of intramolecular connectivity. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 18101–6 (2006).

89. Nitta, R. T., Chu, A. H. & Wong, A. J. Constitutive activity of JNK2 alpha2 is dependent on a unique mechanism of MAPK activation. *J. Biol. Chem.* **283,** 34935–45 (2008).

90. De Nicola, G. F. *et al.* Mechanism and consequence of the autoactivation of p38α mitogen-activated protein kinase promoted by TAB1. *Nat. Struct. Mol. Biol.* **20,** 1182–90 (2013).

91. Pettiford, S. M. & Herbst, R. The MAP-kinase ERK2 is a specific substrate of the protein tyrosine phosphatase HePTP. *Oncogene* **19,** 858–69 (2000).

92. Caunt, C. J. & Keyse, S. M. Dual-specificity MAP kinase phosphatases (MKPs): shaping the outcome of MAP kinase signalling. *FEBS J.* **280,** 489–504 (2013).

93. Akella, R., Moon, T. M. & Goldsmith, E. J. Unique MAP Kinase binding sites. *Biochim. Biophys. Acta* **1784,** 48–55 (2008).

94. Tanoue, T., Adachi, M., Moriguchi, T. & Nishida, E. A conserved docking motif in MAP kinases common to substrates, activators and regulators. *Nat. Cell Biol.* **2,** 110–6 (2000).

95. Nguyen, T., Ruan, Z., Oruganty, K. & Kannan, N. Co-Conserved MAPK Features Couple D-Domain Docking Groove to Distal Allosteric Sites via the C-Terminal Flanking Tail. *PLoS One* **10,** e0119636 (2015).

96. Bhattacharyya, R. P., Reményi, A., Yeh, B. J. & Lim, W. A. Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu. Rev. Biochem.* **75,** 655–80 (2006).

97. Alexandropoulos, K. & Baltimore, D. Coordinate activation of c-Src by SH3- and SH2-binding sites on a novel p130Cas-related protein, Sin. *Genes Dev.* **10,** 1341–55 (1996).

98. Pellicena, P., Stowell, K. R. & Miller, W. T. Enhanced phosphorylation of Src family kinase substrates containing SH2 domain binding sites. *J. Biol. Chem.* **273,** 15325–8 (1998).

99. Lee, K. S., Park, J.-E., Kang, Y. H., Kim, T.-S. & Bang, J. K. Mechanisms underlying Plk1 polo-box domain-mediated biological processes and their physiological significance. *Mol. Cells* **37,** 286–94 (2014).

100. Alessi, D. R. *et al.* The WNK-SPAK/OSR1 pathway: master regulator of cation-chloride cotransporters. *Sci. Signal.* **7,** re3 (2014).

101. Fraser, E. *et al.* Identification of the Axin and Frat binding region of glycogen synthase kinase-3. *J. Biol. Chem.* **277,** 2176–85 (2002).

102. Bax, B. *et al.* The structure of phosphorylated GSK-3beta complexed with a peptide, FRATtide, that inhibits beta-catenin phosphorylation. *Structure* **9,** 1143–52 (2001).

103. Peng, P., Zhao, J., Zhu, Y., Asami, T. & Li, J. A direct docking mechanism for a plant GSK3-like kinase to phosphorylate its substrates. *J. Biol. Chem.* **285,** 24646–53 (2010).

104. Ngo, J. C. K. *et al.* Interplay between SRPK and Clk/Sty kinases in phosphorylation of the splicing factor ASF/SF2 is regulated by a docking motif in ASF/SF2. *Mol. Cell* **20,** 77–89 (2005).

105. Lowe, E. D. *et al.* Specificity determinants of recruitment peptides bound to phospho-CDK2/cyclin A. *Biochemistry* **41,** 15625–34 (2002).

106. Kõivomägi, M. *et al.* Dynamics of Cdk1 substrate specificity during the cell cycle. *Mol. Cell* **42,** 610–23 (2011).

107. Zhang, J., Zhou, B., Zheng, C. & Zhang, Z. A bipartite mechanism for ERK2 recognition by its cognate regulators and substrates. *J. Biol. Chem.* **278,** 29901–12 (2003).

108. Fernandes, N., Bailey, D. E., VanVranken, D. L. & Allbritton, N. L. Use of Docking Peptides to Design Modular Substrates with High Efficiency for Mitogen-Activated Protein Kinase Extracellular Signal-Regulated Kinase. *ACS Chem. Biol.* **2,** 665–673 (2007).

109. Liu, S., Sun, J.-P., Zhou, B. & Zhang, Z.-Y. Structural basis of docking interactions between ERK2 and MAP kinase phosphatase 3. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 5326–31 (2006).

110. Reményi, A., Good, M. C., Bhattacharyya, R. P. & Lim, W. A. The role of docking interactions in mediating signaling input, output, and discrimination in the yeast MAPK network. *Mol. Cell* **20,** 951–62 (2005).

111. Murakami, Y., Tatebayashi, K. & Saito, H. Two adjacent docking sites in the yeast Hog1 mitogen-activated protein (MAP) kinase differentially interact with the Pbs2 MAP kinase kinase and the Ptp2 protein tyrosine phosphatase. *Mol. Cell. Biol.* **28,** 2481–94 (2008).

112. Astigarraga, S. *et al.* A MAPK docking site is critical for downregulation of Capicua by Torso and EGFR RTK signaling. *EMBO J.* **26,** 668–77 (2007).

113. Dinkel, H. *et al.* The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res.* **42,** D259–66 (2014).

114. Heo, Y.-S. *et al.* Structural basis for the selective inhibition of JNK1 by the scaffolding protein JIP1 and SP600125. *EMBO J.* **23,** 2185–95 (2004).

115. Laughlin, J. D. *et al.* Structural mechanisms of allostery and autoinhibition in JNK family kinases. *Structure* **20,** 2174–84 (2012).

116. Chang, C. I., Xu, B., Akella, R., Cobb, M. H. & Goldsmith, E. J. Crystal structures of MAP kinase p38 complexed to the docking sites on its nuclear substrate MEF2A and activator MKK3b. *Mol. Cell* **9,** 1241–9 (2002).

117. Zhou, T., Sun, L., Humphreys, J. & Goldsmith, E. J. Docking Interactions Induce Exposure of Activation Loop in the MAP Kinase ERK2. *Structure* **14,** 1011–1019 (2006).

118. Ma, W. *et al.* Phosphorylation of DCC by ERK2 is facilitated by direct docking of the receptor P1 domain to the kinase. *Structure* **18,** 1502–11 (2010).

119. Pulido, R., Zúñiga, A. & Ullrich, A. PTP-SL and STEP protein tyrosine phosphatases regulate the activation of the extracellular signal-regulated kinases ERK1 and ERK2 by association through a kinase interaction motif. *EMBO J.* **17,** 7337–50 (1998).

120. Mace, P. D. *et al.* Structure of ERK2 bound to PEA-15 reveals a mechanism for rapid release of activated MAPK. *Nat. Commun.* **4,** 1681 (2013).

121. Mészáros, B., Simon, I. & Dosztányi, Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.* **5,** e1000376 (2009).

122. Dosztányi, Z., Mészáros, B. & Simon, I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25,** 2745–6 (2009).

123. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338,** 1027–36 (2004).

124. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.* **35,** W429–32 (2007).

125. Horton, P. *et al.* WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* **35,** W585–7 (2007).

126. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–30 (2014).

127. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252,** 1162–4 (1991).

128. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33,** W382–8 (2005).

129. Sánchez, I. E. *et al.* Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm. *PLoS Comput. Biol.* **4,** e1000052 (2008).

130. Whisenant, T. C. *et al.* Computational prediction and experimental verification of new MAP kinase docking sites and substrates including Gli transcription factors. *PLoS Comput. Biol.* **6,** (2010).

131. Fantz, D. A., Jacobs, D., Glossip, D. & Kornfeld, K. Docking sites on substrate proteins direct extracellular signal-regulated kinase to phosphorylate specific residues. *J. Biol. Chem.* **276,** 27256–65 (2001).

132. Kelkar, N., Gupta, S., Dickens, M. & Davis, R. J. Interaction of a mitogen-activated protein kinase signaling module with the neuronal protein JIP3. *Mol. Cell. Biol.* **20,** 1030–43 (2000).

133. Kallunki, T., Deng, T., Hibi, M. & Karin, M. c-Jun can recruit JNK to phosphorylate dimerization partners via specific docking interactions. *Cell* **87,** 929–39 (1996).

134. Koyano, S. *et al.* A novel Jun N-terminal kinase (JNK)-binding protein that enhances the activation of JNK by MEK kinase 1 and TGF-beta-activated kinase 1. *FEBS Lett.* **457,** 385–8 (1999).

135. Christova, Y., Adrain, C., Bambrough, P., Ibrahim, A. & Freeman, M. Mammalian iRhoms have distinct physiological functions including an essential role in TACE regulation. *EMBO Rep.* **14,** 884–90 (2013).

136. Maretzky, T. *et al.* iRhom2 controls the substrate selectivity of stimulated ADAM17-dependent ectodomain shedding. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 11433–8 (2013).

137. Theodosiou, A., Smith, A., Gillieron, C., Arkinstall, S. & Ashworth, A. MKP5, a new member of the MAP kinase phosphatase family, which selectively dephosphorylates stress-activated kinases. *Oncogene* **18,** 6981–8 (1999).

138. Zhang, Y.-Y., Wu, J.-W. & Wang, Z.-X. A distinct interaction mode revealed by the crystal structure of the kinase p38α with the MAPK binding domain of the phosphatase MKP5. *Sci. Signal.* **4,** ra88 (2011).

139. Gdalyahu, A. *et al.* DCX, a new mediator of the JNK pathway. *EMBO J.* **23,** 823–32 (2004).

140. Okamoto, M. & Südhof, T. C. Mints, Munc18-interacting proteins in synaptic vesicle exocytosis. *J. Biol. Chem.* **272,** 31459–64 (1997).

141. Kragelj, J. *et al.* Structure and dynamics of the MKK7-JNK signaling complex. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 3409–14 (2015).

142. Dajas-Bailador, F., Jones, E. V & Whitmarsh, A. J. The JIP1 scaffold protein regulates axonal development in cortical neurons. *Curr. Biol.* **18,** 221–6 (2008).

143. Bogoyevitch, M. A. *et al.* WD40-repeat protein 62 is a JNK-phosphorylated spindle pole protein required for spindle maintenance and timely mitotic progression. *J. Cell Sci.* **125,** 5096–5109 (2012).

144. Kim, J. G. *et al.* Myelin transcription factor 1 (Myt1) of the oligodendrocyte lineage, along with a closely related CCHC zinc finger, is expressed in developing neurons in the mammalian central nervous system. *J. Neurosci. Res.* **50,** 272–90 (1997).

145. Lamba, D. A., Hayes, S., Karl, M. O. & Reh, T. Baf60c is a component of the neural progenitor-specific BAF complex in developing retina. *Dev. Dyn.* **237,** 3016–23 (2008).

146. Lu, Q. R., Cai, L., Rowitch, D., Cepko, C. L. & Stiles, C. D. Ectopic expression of Olig1 promotes oligodendrocyte formation and reduces neuronal survival in developing mouse cortex. *Nat. Neurosci.* **4,** 973–4 (2001).

147. Thauvin-Robinet, C. *et al.* The oral-facial-digital syndrome gene C2CD3 encodes a positive regulator of centriole elongation. *Nat. Genet.* **46,** 905–11 (2014).

148. Heydet, D. *et al.* A truncating mutation of Alms1 reduces the number of hypothalamic neuronal cilia in obese mice. *Dev. Neurobiol.* **73,** 1–13 (2013).

149. Yau, K. W. *et al.* Microtubule minus-end binding protein CAMSAP2 controls axon specification and dendrite development. *Neuron* **82,** 1058–73 (2014).

150. Horie, M. *et al.* Disruption of actin-binding domain-containing Dystonin protein causes dystonia musculorum in mice. *Eur. J. Neurosci.* **40,** 3458–71 (2014).

151. Watabe-Uchida, M., John, K. A., Janas, J. A., Newey, S. E. & Van Aelst, L. The Rac activator DOCK7 regulates neuronal polarity through local phosphorylation of stathmin/Op18. *Neuron* **51,** 727–39 (2006).

152. Martínez-López, M. J. *et al.* Mouse neuron navigator 1, a novel microtubule-associated protein involved in neuronal migration. *Mol. Cell. Neurosci.* **28,** 599–612 (2005).

153. Simon-Areces, J. *et al.* Formin1 mediates the induction of dendritogenesis and synaptogenesis by neurogenin3 in mouse hippocampal neurons. *PLoS One* **6,** e21825 (2011).

154. Deller, T. *et al.* Synaptopodin-deficient mice lack a spine apparatus and show deficits in synaptic plasticity. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 10494–9 (2003).

155. Oliva, A. A., Atkins, C. M., Copenagle, L. & Banker, G. A. Activated c-Jun N-terminal kinase is required for axon formation. *J. Neurosci.* **26,** 9462–70 (2006).

156. Myers, A. K., Meechan, D. W., Adney, D. R. & Tucker, E. S. Cortical interneurons require Jnk1 to enter and navigate the developing cerebral cortex. *J. Neurosci.* **34,** 7787–801 (2014).

157. Hirosumi, J. *et al.* A central role for JNK in obesity and insulin resistance. *Nature* **420,** 333–6 (2002).

158. Li, L. *et al.* IG20/MADD plays a critical role in glucose-induced insulin secretion. *Diabetes* **63,** 1612–23 (2014).

159. Lee, Y. H., Giraud, J., Davis, R. J. & White, M. F. c-Jun N-terminal kinase (JNK) mediates feedback inhibition of the insulin signaling cascade. *J. Biol. Chem.* **278,** 2896–902 (2003).

160. Olson, B. L. *et al.* SCFCdc4 acts antagonistically to the PGC-1alpha transcriptional coactivator by targeting it for ubiquitin-mediated proteolysis. *Genes Dev.* **22,** 252–64 (2008).

161. Mihaylova, M. M. & Shaw, R. J. The AMPK signalling pathway coordinates cell growth, autophagy and metabolism. *Nat. Cell Biol.* **13,** 1016–23 (2011).

162. Costanzo-Garvey, D. L. *et al.* KSR2 is an essential regulator of AMP kinase, energy expenditure, and insulin sensitivity. *Cell Metab.* **10,** 366–78 (2009).

163. Lang, T. *et al.* Molecular Cloning, Genomic Organization, and Mapping of PRKAG2, a Heart Abundant γ2 Subunit of 5′-AMP-Activated Protein Kinase, to Human Chromosome 7q36. *Genomics* **70,** 258–263 (2000).

164. Pearce, L. R. *et al.* KSR2 mutations are associated with obesity, insulin resistance, and impaired cellular fuel oxidation. *Cell* **155,** 765–77 (2013).

165. Jensen, L. J. *et al.* eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **36,** D250–4 (2008).

166. Berglund, A.-C., Sjölund, E., Ostlund, G. & Sonnhammer, E. L. L. InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res.* **36,** D263–6 (2008).

167. Goldman, A. *et al.* The calcineurin signaling network evolves via conserved kinase-phosphatase modules that transcend substrate identity. *Mol. Cell* **55,** 422–35 (2014).

168. Ohno, S. Patterns in genome evolution. *Curr. Opin. Genet. Dev.* **3,** 911–4 (1993).

169. Gordon, E. A. *et al.* Combining docking site and phosphosite predictions to find new substrates: identification of smoothelin-like-2 (SMTNL2) as a c-Jun N-terminal kinase (JNK) substrate. *Cell. Signal.* **25,** 2518–29 (2013).

170. Wolf, A. *et al.* MAPK-induced Gab1 translocation to the plasma membrane depends on a regulated intramolecular switch. *Cell. Signal.* **27,** 340–52 (2015).

171. Eulenfeld, R. & Schaper, F. A new mechanism for the regulation of Gab1 recruitment to the plasma membrane. *J. Cell Sci.* **122,** 55–64 (2009).

172. Sanders, M. A., Ampasala, D. & Basson, M. D. DOCK5 and DOCK1 regulate Caco-2 intestinal epithelial cell spreading and migration on collagen IV. *J. Biol. Chem.* **284,** 27–35 (2009).

173. Podkowa, M. *et al.* Microtubule stabilization by bone morphogenetic protein receptor-mediated scaffolding of c-Jun N-terminal kinase promotes dendrite formation. *Mol. Cell. Biol.* **30,** 2241–50 (2010).

174. Chen, W.-K., Yeap, Y. Y. C. & Bogoyevitch, M. A. The JNK1/JNK3 interactome - Contributions by the JNK3 unique N-terminus and JNK common docking site residues. *Biochem. Biophys. Res. Commun.* **453,** 576–81 (2014).

175. Ho, D. T., Bardwell, A. J., Grewal, S., Iverson, C. & Bardwell, L. Interacting JNK-docking sites in MKK7 promote binding and activation of JNK mitogen-activated protein kinases. *J. Biol. Chem.* **281,** 13169–79 (2006).

176. Nagadoi, A. *et al.* Solution structure of the transactivation domain of ATF-2 comprising a zinc finger-like subdomain and a flexible subdomain. *J. Mol. Biol.* **287,** 593–607 (1999).

177. Chang, Y. S. *et al.* Stapled α-helical peptide drug development: a potent dual inhibitor of MDM2 and MDMX for p53-dependent cancer therapy. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E3445–54 (2013).

178. Farooq, A. *et al.* Solution structure of ERK2 binding domain of MAPK phosphatase MKP-3: structural insights into MKP-3 activation by ERK2. *Mol. Cell* **7,** 387–99 (2001).

179. Slack, D. N., Seternes, O. M., Gabrielsen, M. & Keyse, S. M. Distinct binding determinants for ERK2/p38alpha and JNK map kinases mediate catalytic activation and substrate selectivity of map kinase phosphatase-1. *J. Biol. Chem.* **276,** 16491–500 (2001).

180. Taru, H. & Suzuki, T. Facilitation of stress-induced phosphorylation of beta-amyloid precursor protein family members by X11-like/Mint2 protein. *J. Biol. Chem.* **279,** 21628–36 (2004).

181. Kosako, H. *et al.* Phosphoproteomics reveals new ERK MAP kinase targets and links ERK to nucleoporin-mediated nuclear transport. *Nat. Struct. Mol. Biol.* **16,** 1026–35 (2009).

182. Carlson, S. M. *et al.* Large-scale discovery of ERK2 substrates identifies ERK-mediated transcriptional regulation by ETV3. *Sci. Signal.* **4,** rs11 (2011).

183. Courcelles, M. *et al.* Phosphoproteome dynamics reveal novel ERK1/2 MAP kinase substrates with broad spectrum of functions. *Mol. Syst. Biol.* **9,** 669 (2013).

184. Bandyopadhyay, S. *et al.* A human MAP kinase interactome. *Nat. Methods* **7,** 801–5 (2010).

185. Park, S., Uesugi, M. & Verdine, G. L. A second calcineurin binding site on the NFAT regulatory domain. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 7130–5 (2000).

186. Iakoucheva, L. M. *et al.* The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32,** 1037–49 (2004).

187. Nishi, H., Fong, J. H., Chang, C., Teichmann, S. A. & Panchenko, A. R. Regulation of protein-protein binding by coupling between phosphorylation and intrinsic disorder: analysis of human protein complexes. *Mol. Biosyst.* **9,** 1620–6 (2013).

188. Wu, W., de Folter, S., Shen, X., Zhang, W. & Tao, S. Vertebrate paralogous MEF2 genes: origin, conservation, and evolution. *PLoS One* **6,** e17334 (2011).

189. Inuzuka, H. *et al.* SCF(FBW7) regulates cellular apoptosis by targeting MCL1 for ubiquitylation and destruction. *Nature* **471,** 104–9 (2011).

190. Bao, M. Z., Shock, T. R. & Madhani, H. D. Multisite phosphorylation of the Saccharomyces cerevisiae filamentous growth regulator Tec1 is required for its recognition by the E3 ubiquitin ligase adaptor Cdc4 and its subsequent destruction in vivo. *Eukaryot. Cell* **9,** 31–6 (2010).

191. Wei, W., Jin, J., Schlisio, S., Harper, J. W. & Kaelin, W. G. The v-Jun point mutation allows c-Jun to escape GSK3-dependent recognition and destruction by the Fbw7 ubiquitin ligase. *Cancer Cell* **8,** 25–33 (2005).

192. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19,** 327–35 (2009).

SUPPLEMENTARY MATERIALS

List of supplementary materials (available in electronic format only, due to their size)

Supplementary tables

Supplementary figures